HCF-1 inhibits SKN-1 to modulate stress resistance but not lifespan in *Caenorhabditis elegans*

and

Determination of enrichment regions for H3K27me3 and other low-signal, high-noise ChIP-seq data

Honors Thesis
Presented to the College of Agriculture and Life Sciences
Cornell University
in Partial Fulfillment of the Requirements for the
Biology Honors Program

by
Colette Lafontaine Picard
May 2012

Supervisor Siu Sylvia Lee

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

I would like to thank my research advisor, Dr. Siu Sylvia Lee, for her invaluable help and support over the last three years. I would also like to thank Dr. Gizem Rizki, with whom I did much of the first project presented in this paper and who taught me so much of what I know about doing research in a lab. I also thank Ella Chang and everyone else in the Lee lab for their help and suggestions throughout these projects. Outside the Lee lab, both Drs. Jim Booth and Haim Bar provided great statistical advice, and Dr. Adam Siepel and Charles Danko were a source of very useful advice and insights regarding peak-calling algorithms. In addition, I would like to thank the Howard Hughes Program for providing funding for me to stay and work in the lab during the summer of 2010. Finally, I would like to thank my father Robert for his invaluable tech support and debugging skills (both in code and in thoughts), and my mother Francine for all her advice, help, proofreading and cheering on.

INTRODUCTION

This thesis is comprised of two projects. The first project (Chapter 1) seeks to determine the role of *Caenorhabditis elegans* SKN-1 in regulating HCF-1 mediated lifespan and stress resistance. The second project (Chapter 2) investigates peak-calling methods used to analyze ChIP-seq data with broad regions of enrichment.

Chapter 1: HCF-1 inhibits SKN-1 to modulate stress resistance but not lifespan in *Caenorhabditis elegans*

ABSTRACT

Caenorhabditis elegans host cell factor-1 (HCF-1) is an evolutionarily conserved longevity determinant. HCF-1 modulates both lifespan and stress resistance by inhibiting the *C. elegans* homolog of the mammalian FOXO transcription factors, DAF-16. DAF-16 is another well-characterized and conserved longevity determinant best known as the primary target of the insulin/IGF-1-like signaling (IIS) pathway. However, the involvement of other components in HCF-1 mediated longevity and stress resistance has not yet been characterized. We show that SKN-1, the *C. elegans* homolog of the mammalian Nrf proteins and a major orchestrator of the phase II detoxification response that defends against oxidative stress, is regulated by HCF-1 to modulate oxidative stress resistance but not lifespan. We find that HCF-1 acts to suppress SKN-1 activity when the worm is exposed to oxidative stress, and that SKN-1 is required for HCF-1 mediated stress resistance. However, SKN-1 is not required for HCF-1 mediated longevity. Our findings imply a novel regulatory relationship between HCF-1 and SKN-1 that is revealed only in the presence of oxidative stress.

INTRODUCTION[1]

Cells have evolved robust mechanisms to reduce the damage caused by oxidative stress. Cells are exposed to Reactive Oxygen Species (ROS) as a natural by-product of cellular respiration produced by mitochondria, as well as from the environment. When exposed to oxidative stress, eukaryotic cells activate several signaling cascades that ultimately cause the expression of various enzymes whose role is to degrade the ROS (An and Blackwell 2003). When ROS levels become sufficiently high, or the efficiency of the oxidative stress resistance pathways is compromised, ROS can cause significant damage to cells, leading to the development of various diseases including degenerative disorders, diabetes and cancer (Reuter *et al*. 2010; Rains *et al*. 2011; Oliveira *et al*. 2009). The pathways associated with oxidative stress resistance are well conserved, and in mammals involve both the forkhead box O (FOXO) and NF-E2-related (Nrf) families of transcription factors.

The *C. elegans* Nrf ortholog SKN-1 has been shown to have a similar role as Nrf in mammals and is responsible for coordinating the response to oxidative stress (An and Blackwell 2003). The FOXO ortholog in worms, DAF-16, is a well-characterized transcription factor known to play a major role in coordinating diverse cellular processes from aging to metabolism and stress resistance (Kenyon *et al*. 1993; Ogg *et al*. 1997; Lee *et al*. 2003; Amrit *et al*. 2010). Notably, DAF-16 is the primary downstream target of the insulin/IGF-1-like signaling (IIS) pathway (Kenyon *et al*. 1993). IIS inhibits DAF-16 by phosphorylation that causes cytosolic sequestration. Knocking out various components of the IIS pathway causes a robust and dramatic lifespan extension, as well as increased resistance to various stressors including oxidative stress (Kenyon *et al*. 1993, Honda and Honda 1999).

In the nucleus, *C. elegans* host cell factor 1 (HCF-1) acts as another suppressor of DAF-16 activity, and *hcf-1* mutants have an increased lifespan that is also dependent on DAF-16 (Li *et al*. 2008). In addition, *hcf-1* mutants display increased resistance to oxidative stress in a manner also dependent on DAF-16 (Li *et al*. 2008). However, the

[1] While much of this project was done in collaboration with Gizem Rizki, who performed many experiments similar to those presented here, all data presented and analyzed in the present paper were generated solely by Colette L. Picard.

involvement of additional components in HCF-1-mediated lifespan and oxidative stress resistance has not yet been characterized.

It has been shown recently that IIS signaling inhibits SKN-1 in parallel to DAF-16, phosphorylating both transcription factors to prevent their translocation into the nucleus and thereby preventing access to their target genes (Tullet *et al*. 2008). We reasoned that if HCF-1 acts to oppose longevity and stress resistance by suppressing DAF-16 activity, it might similarly suppress SKN-1, which has similar roles to DAF-16 with respect to longevity and stress resistance, and is also suppressed by IIS. Our analyses and results show that the increased stress resistance, but not lifespan, phenotype of *hcf-1* mutant worms is dependent on the presence of SKN-1 (Rizki *et al*, manuscript in revision). We conclude that HCF-1 inhibits both DAF-16 and SKN-1, but that while it affects DAF-16 to modulate both lifespan and stress resistance, it affects SKN-1 only with respect to stress resistance (model shown in Fig. 1).

RESULTS

Lifespan epistasis assays

To determine the interaction between SKN-1 and HCF-1 in regulating longevity, my co-authors and I treated *hcf-1* loss-of-function mutants with RNAi targeting *skn-1* (Rizki *et al*, Aging Cell, *in press*). We then examined lifespan to see whether the *skn-1* RNAi could suppress the increased lifespan phenotype of *hcf-1* mutant worms (Fig. 2A, Table 1A). Our results show that *skn-1* RNAi causes a slight decrease in lifespan in both wild-type and *hcf-1* backgrounds, but the *hcf-1* mutants treated with *skn-1* RNAi continue to live significantly longer than wild-type worms on *skn-1* RNAi. This suggests that HCF-1 can modulate lifespan even when SKN-1 is largely inactivated. Data shown in Fig. 2A are pooled from five separate experiments, two of which included the *daf-16* RNAi and three of which did not. Consistent with published results (Li *et al*. 2008), the *daf-16* RNAi used as a positive control suppressed the *hcf-1* lifespan phenotype almost completely.

Tullet *et al*. (2008) reported that knockdown of *skn-1* was only able to decrease the increased lifespan conferred by mutation in the insulin receptor *daf-2* when performed in the RNAi-sensitive *rrf-3* background. To confirm that the RNAi efficiency was not an

issue here, we repeated our lifespan experiments in the RNAi-sensitive *rrf-3* background (see Fig. 2B, Table 1B). The results were again consistent with our earlier conclusion that HCF-1 does not depend on SKN-1 to modulate lifespan in *C. elegans*. Data for experiments in *rrf-3* background are pooled from three separate experiments, one of which included the *daf-16* RNAi and two of which did not. Again, *daf-16* fully suppressed the *hcf-1* mutant lifespan. Taken together, these results suggest that SKN-1 is not required for HCF-1-mediated longevity.

SKN-1 target gene GCS-1::GFP is induced in *hcf-1* background in the absence of oxidative stress

We next considered whether HCF-1 affects the transcriptional activity of SKN-1. First, I attempted to determine whether the presence of HCF-1 inhibited the induction of a known SKN-1 target gene. γ-glutamyl cysteine synthetase (*gcs-1*) is a Phase II detoxification gene directly induced by SKN-1 in the intestine, but normally only under stress conditions (An and Blackwell 2003). Interestingly, knocking down *hcf-1* induced GCS-1::GFP expression in the intestine in the absence of stress (Fig. 3A, Table 2). Similar experiments conducted by Gizem Rizki for two other SKN-1 targets, glutathione S-transferases *gst-4* and *gst-7*, also showed that *hcf-1* RNAi knockdown leads to the upregulation of these genes in the absence of stress (Rizki *et al*, Aging Cell, *in press*). Notably, at least in the case of GCS-1, this induction does require SKN-1, since the GCS-1::GFP induction caused by *hcf-1* RNAi was completely suppressed in the *skn-1* mutant background (Fig. 3A, Table 2). These results suggest that under basal conditions, HCF-1 suppresses the expression of SKN-1 target genes.

Intestinal SKN-1::GFP nuclear localization is unaffected in *hcf-1* background in the absence of oxidative stress

After observing changes in the expression of SKN-1 target genes in response to *hcf-1* knockout under basal conditions, we attempted to determine the mechanism by which SKN-1 and HCF-1 were interacting to cause these changes. Normally, SKN-1 is induced into intestinal nuclei only in response to stress, allowing it to activate its target genes and coordinate the stress response (An and Blackwell 2003). My co-authors and I

reasoned that one mechanism through which HCF-1 could inhibit SKN-1 would be by preventing SKN-1 translocation into the nucleus. If this were the case, we would expect to see greater SKN-1 nuclear localization in *hcf-1* mutants than in wild-type, which would allow SKN-1 greater access to target genes such as GCS-1 and cause the upregulation of GCS-1 observed earlier. To test these predictions, I treated worms expressing a transgenic form of SKN-1 fused with GFP (SKN-1::GFP) with control, *akt-1/akt-2* and *hcf-1* RNAi. I then scored the degree of SKN-1::GFP nuclear localization as shown in Fig. 4. AKT-1 and AKT-2 are both members of the insulin signaling pathway, and therefore knocking them down induced SKN-1::GFP into the nucleus as expected from Tullet *et al*. (2008). However, knockdown of *hcf-1* failed to alter SKN-1::GFP nuclear localization compared to control RNAi (Fig. 3B, Table 2), suggesting that HCF-1 does not affect SKN-1 nuclear localization under basal conditions.

SKN-1::GFP is further translocated into the nucleus in *hcf-1* background in the presence of oxidative stress

Interestingly, while SKN-1 and HCF-1 do not appear to interact to modulate lifespan, our data suggest that the increased oxidative stress phenotype of *hcf-1* mutants does depend on SKN-1. Rizki *et al*. (Aging Cell, *in press*) performed experiments measuring the oxidative stress resistance phenotype of *hcf-1* mutants when subjected to *skn-1* RNAi knockdown. Rizki *et al*. placed worms on *tert*-Butyl Hydroperoxide (*t*-BOOH), paraquat or Sodium Arsenite (NaAs), all commonly used to induce oxidative stress in worms, and showed that the elevated stress resistance phenotype of *hcf-1* mutants was suppressed by *skn-1*. While earlier results suggested that HCF-1 does not affect SKN-1 nuclear localization under basal conditions, I performed further fluorescence microscopy assays to show that *skn-1::gfp* is more highly induced into intestinal nuclei in *hcf-1* mutant worms than wild-type when exposed to stress (Fig. 2C and 2D).

In particular, to determine how HCF-1 affects SKN-1 nuclear localization under stress conditions, I crossed SKN-1::GFP worms into the *hcf-1(pk924)* background (see methods). I then subjected wild-type or *hcf-1* young adult/early gravid adult worms to either 10mM Sodium Arsenite or 2mM *t*-BOOH. For the NaAs assay (Fig. 3C, Table 3),

worms were grown on stock plates until young adult/ early gravid adult before being transferred onto stock plates containing either 10mM NaAs or M9 buffer, the latter acting as a vehicle control since the NaAs was dissolved in M9. In this experiment, SKN-1::GFP was highly induced into the intestinal nuclei of NaAs treated worms even in the presence of HCF-1, but this induction was even higher in *hcf-1* mutant worms. Unexpectedly, the results suggest that the M9 buffer also causes a stress response in worms since exposure caused mild induction of SKN-1::GFP even in the presence of HCF-1. The induction of SKN-1 when exposed to M9 was even more dramatic in *hcf-1* mutant background. Notably, previous experiments (see Fig. 3B) did not show a change in SKN-1 localization in response to *hcf-1* knockdown in the absence of stress. However, if M9 exposure does in fact represent a mild source of stress, then the further induction of SKN-1 in *hcf-1* background is not surprising and is consistent with our observations that HCF-1 affects SKN-1 nuclear localization only in the presence of stress.

For the assays involving t-BOOH (Fig. 3D, Table 3), worms were similarly grown to the young adult/ early gravid adult stage before being transferred to plates containing either t-BOOH or decane, which is used to dissolve the t-BOOH and acts here as a vehicle control. Not shown are the water control results, where worms were transferred to plates onto which I had added only water. Worms on the water control plates predictably failed to show any nuclear translocation of SKN-1::GFP. The results of this experiment also showed increase in SKN-1 nuclear accumulation in response to stress in the *hcf-1* background when compared to wild type. In these experiments, unlike in the NaAs assay, I did not see further induction of SKN-1::GFP into the nucleus in *hcf-1* mutants under the control (decane) condition, suggesting that decane, unlike M9 buffer, does not activate the *C. elegans* stress response pathways. These results are consistent with our previous observations (Fig. 3B) that HCF-1 acts to suppress SKN-1 nuclear translocation only under oxidative stress conditions. Taken together, my results suggest that HCF-1 regulates SKN-1 nuclear localization only under oxidative stress conditions, such as NaAs or *t*-BOOH, and that unexpectedly, exposure to M9 buffer seems to also activate the stress response. HCF-1 also suppresses the induction of known SKN-1 targets *gcs-1, gst-4* and *gst-7*, all of which are normally induced only in response to stress, but which are induced in the absence of stress in *hcf-1* mutants. It is possible that

SKN-1 is mildly activated in the *hcf-1* mutants in the absence of stress, but that mild activation is not reflected as nuclear translocation under the experimental conditions I tested.

Microarray analysis of *hcf-1(-)* and *skn-1(+)* worms

I also obtained microarray data comparing the mRNA levels of wild-type worms treated with control or *skn-1* RNAi (indicated as *skn-1(+)* below) (Oliveira et al. 2009), and comparing *hcf-1(pk924)* null mutant worms to wild-type (indicated as *hcf-1(-)* below) (Rizki *et al*, Aging Cell, *in press*). I used the microarray data analysis tool Significance Analysis of Microarrays (SAM) to identify genes that are similarly expressed in *hcf-1(-)* and *skn-1(+)*, as well as genes that are oppositely expressed and genes that are affected only by *hcf-1* (Tusher *et al*. 2001). Hierarchical clustering was performed on all genes flagged in at least one of these SAM runs just described. Results from these analyses are shown in Fig. 5. There were a total of 190 genes similarly affected in *hcf-1(-)* and *skn-1(+)*, 206 genes oppositely affected, and 640 genes affected only by *hcf-1(-)*. The high number of genes affected only by HCF-1 suggests that HCF-1 has many roles beside those that are related to SKN-1, which is consistent with our observation that SKN-1 seems to be required for the *hcf-1* oxidative stress resistance phenotype but not the longevity or other *hcf-1* phenotypes. The lists of genes identified by SAM and hierarchical clustering were submitted for functional annotation clustering and Gene Ontological analysis using DAVID (Huang *et al*. 2009). Genes from clusters (A) and (E) in Fig. 5 were combined into a single group for DAVID analysis, as were genes from clusters (B) and (D). The results of the GO-term analysis are summarized in Table 4. Strikingly, there was a single dramatically enriched cluster in the group of genes similarly affected in *hcf-1(-)* and *skn-1(+)*, centering around the Phase II Detoxification response, which is mediated by SKN-1 (An and Blackwell 2003). This would suggest that HCF-1 and SKN-1 work together to regulate the expression of genes related to this detoxification response, and is again consistent with the idea that SKN-1 and HCF-1 interact only with respect to the stress response. GO-terms associated with aging appeared in all three groups, but were more highly enriched relative to other GO-terms in the 'opposite' and '*hcf-1* only' clusters (appearing in the top 5 in both cases). This result further suggests that HCF-1's longevity mediating role is independent of SKN-1.

DISCUSSION

Mammalian HCF-1 is thought to regulate gene expression through several different mechanisms, including by regulating other transcription factors and chromatin factors, and by assembling protein complexes for context-dependent gene regulation (Li *et al*. 2008). Given the complex mechanisms by which HCF-1 regulates gene expression in mammals, and the high degree of structural and functional conservation between the nematode and mammalian HCF-1 proteins (Liu et al. 1999), it is likely that *C. elegans* HCF-1 interacts with many other factors in order to bring about transcriptional changes. In this paper, we have identified *C. elegans* SKN-1 as a novel contributor to HCF-1 mediated oxidative stress resistance, but not HCF-1 mediated longevity. I showed that the lifespan of *hcf-1* mutant worms is not dependent on the presence of SKN-1, but results from Rizki *et al*. (Aging Cell, *in press*) showed that the stress resistance phenotype of *hcf-1* mutants was SKN-1 dependent. Analysis of microarray data comparing *hcf-1(-)* mRNA levels to *skn-1(+)* mRNA levels further showed that genes that are similarly changed in *hcf-1(-)* and *skn-1(+)* are dramatically enriched for genes that are associated with oxidative stress response pathways, while genes oppositely changed in the two groups, or changed only in *hcf-1(-)*, were more enriched for aging-related genes.

The fluorescence microscopy assays performed in this paper give some insight into how HCF-1 might suppress SKN-1 activity. SKN-1 normally accumulates in the nucleus of intestinal cells only in response to oxidative stress, where it acts as a primary activator of the major oxidative stress response pathways (An and Blackwell 2003). Our findings support this and further suggest that under basal conditions, SKN-1 nuclear localization is unaffected by *hcf-1*. Interestingly, however, we see a dramatic impact on SKN-1 nuclear localization by *hcf-1* in oxidative stress conditions. From these data, we cannot determine whether the increase in SKN-1 localization is due to increased import of SKN-1 into the nucleus, decreased export, decreased SKN-1 degradation or an overall increase in SKN-1 expression. It is therefore difficult to determine how HCF-1 can regulate SKN-1 nuclear localization only under specific conditions such as oxidative stress exposure. HCF-1 is known to bind directly with DAF-16 to prevent DAF-16 localization to its target genes (Li *et al*. 2008). It is unknown whether or not HCF-1

similarly binds with SKN-1. It is possible that HCF-1 represses SKN-1 activity by forming a complex with it that prevents SKN-1 access to its target genes. Alternatively, it is possible that HCF-1, which in both worms and mammals has been shown to regulate other transcription factors by recruiting chromatin modifiers to promoters (Lee *et al*. 2007), alters the expression of SKN-1 target genes in this way. Both explanations, however, fail to explain the changes in SKN-1 nuclear localization observed in this paper. Further research is needed to fully characterize the method through which HCF-1 mediates SKN-1 nuclear localization and target gene expression in the context of oxidative stress. It is also unknown whether HCF-1 and SKN-1 interact to regulate other stress response pathways. Blackwell and An (2003) showed a dramatic increase in SKN-1 nuclear localization in response to heat, suggesting that SKN-1 plays a role in regulating the heat shock response. It would be of interest to determine whether the SKN-1 mediated heat shock response is also regulated by HCF-1, and whether the same pattern of nuclear localization regulation by HCF-1 appears.

One point of interest is that our findings appear to draw a line between longevity and oxidative stress resistance, suggesting that these are two largely separate pathways when mediated by HCF-1. Longevity and increased stress resistance are usually observed together, since increased resistance to oxidative stress seems to have a protective effect on lifespan even when organisms are not exposed to stress (Hyun *et al*. 2007). Recent studies have shown that, contrary to the predictions of Harman's free radical theory of aging, increased exposure to mild oxidative stress, which activates the body's natural stress response pathways, can have a beneficial impact on longevity (Ristow and Zarse 2010; Hyun *et al*. 2007; Harman 1956). Artificial activation of these pathways through mutations such as *daf-2*, a component of the IIS pathway, has also been shown to confer not only increased stress resistance but also increased longevity (Kenyon *et al*. 1993; Tullet *et al*. 2008). Defects in these same pathways often cause phenotypes similar to premature aging (Hyun *et al*. 2007). The question remains how much of the increased longevity observed in *daf-2* and in other mutants is due to increased stress resistance. Our results suggest that the increased stress resistance conferred by *hcf-1* mutation through SKN-1 activation has a minimal impact on the longevity of *hcf-1* mutants. It is possible that the stress pathways regulated by SKN-1 in particular do not

impact longevity, since lifespan assays in both this paper and An and Blackwell (2003) show little difference between the longevity of *skn-1* mutants when compared to wild-type. In addition, Li *et al*. 2008 showed that DAF-16 is required for both *hcf-1* longevity and stress resistance, suggesting that the stress resistance pathways activated by DAF-16, but not those activated by SKN-1, in *hcf-1* mutants confer increased longevity. An explanation for why the DAF-16 stress resistance pathways would impact longevity, while the SKN-1 pathways do not, remains elusive. Further studies to determine the extent with which different stress response pathways contribute to longevity could be of interest. However, it is important to keep in mind that the results presented in this paper consist only of lifespan experiments conducted using *skn-1* RNAi treated *hcf-1* mutant worms, rather than using *hcf-1; skn-1* double mutants as is typically done to evaluate the epistatic relationship of two genes. Therefore we cannot conclude that SKN-1 is not required for *hcf-1* mediated lifespan. Unfortunately, *skn-1* and *hcf-1* are both on chromosome IV in the worm, somewhat close together. This, in addition to the fact that the *skn-1* null mutation is lethal (homozygotes survive due to maternal *skn-1* but are themselves sterile) and needs to be balanced, make generating double mutants to conduct this experiment quite challenging. Therefore, constructing the *hcf-1 skn-1* double mutants was not attempted here. However, the reverse experiment, in which *skn-1* mutants are treated with either *hcf-1* or control RNAi, could be performed to confirm the results presented in this paper. If confirmed, the dissociation between *hcf-1* longevity and SKN-1 mediated stress resistance would be an interesting departure from what is normally observed, and could provide important insights into how some forms of oxidative stress resistance confer increased longevity.

In summary, the interaction of HCF-1 and SKN-1 represents an important mechanism by which cells defend themselves from oxidative stressors. In addition, the high degree of conservation between HCF-1 and SKN-1 in worms and mammals suggests that a relationship between mammalian HCF-1 and Nrf factors may also exist, and may represent an important pathway in the regulation of oxidative-stress induced disease including neurodegenerative disorders and cancer.

MATERIALS AND METHODS

C. elegans strains

Strains are maintained on stock plates of *e. coli* OP50 bacteria seeded onto Nematode Growth Medium (NGM) plates at 16°C. Strains used are: N2 wild type, *hcf-1(pk924)*, *daf-16(mgDf47)*, IU162.1 *daf-16(mgDf47);hcf-1(pk924)*, LD1002 *Ex[gcs-1::gfp pRF4 rol-6]* (An and Blackwell 2003), LD1004 *Ex[skn-1b/c::gfp pRF4 rol-6]* (An and Blackwell 2003), IU408.1 and IU408.2 *hcf-1 (pk924) Ex[skn-1b/c::gfp pRF4 rol-6]* (two isolates), *rrf-3(pk1426), rrf-3(pk1426);hcf-1(pk924)*.

Lifespan experiments

Well fed gravid worms were allowed to lay eggs at room temperature overnight (~3 worms/plate for strains with normal or slightly reduced fertility – N2, *hcf-1*, *rrf-3* – and ~10/plate for strains with limited fertility – *hcf-1;rrf-3*). The progeny were grown at 25°C until young adult/early gravid stage (called Day 0) and then 35 progeny per plate were transferred onto new plates containing fresh FUDR. Worms were transferred again onto fresh food and FUDR on days 2, 4 and 8. Adult worms were scored every other day, and worms that failed to move after being gently prodded several times by a platinum wire pick were scored as dead. Worms that crawled onto the wall of the plate or had a large protruding or burst vulva were censored. SPSS software was used to perform the data analysis and generate the survival curves. The Kaplan-Meier log rank test was used to determine whether independent experiments were significantly different. These results are shown in Table 1A and 1B.

Crosses

To create *hcf-1(pk924)* worms expressing SKN-1::GFP (LD1004), 15 male *hcf-1* worms were placed on a mating plate with 5 wild-type L4 hermaphrodites and allowed to mate at 20°C for two days. After mating, each hermaphrodite was transferred onto an individual plate. Progeny were monitored for the 1:1 male to hermaphrodite ratio that is expected if the worm was successfully mated. From the plates that showed successful mating, a total of 4 F1 worms were placed on individual plates and allowed to self-fertilize, also at 20°C. After about 2-3 days (F2 progeny at L4 stage), 12 F2 worms per F1 plate (48 in total)

were placed on individual plates and also allowed to self-fertilize. Plates on which all progeny displayed the GFP phenotype were separated out and genotyped to identify *hcf-1(pk924)* homozygous worms containing the transgene.

Nuclear Localization Experiments

All of the nuclear localization experiments were performed blind – e.g. the experimenter scoring the nuclear localization of the worms did not know which RNAi or stress condition was being observed. *p*-values shown in Tables 2 and 3 were calculated using a chi-squared test.

Nuclear Localization Experiments - Scoring

Worms were viewed under a fluorescence microscope with a GFP filter. Worms scored as 'high' showed a strong GFP signal in all intestinal nuclei. Medium scores were used to indicate that nuclear SKN-1::GFP signal was high anteriorly, posteriorly or both, but did not appear midway through the intestine. A worm was also scored as medium if it displayed a weak SKN-1::GFP signal in all intestinal nuclei. Low scores were used to indicate no nuclear localization, i.e. no GFP was evident beyond the ASI neurons. See Fig. 4 for illustration of SKN-1::GFP scoring scheme.

Nuclear Localization Experiments - hcf-1 RNAi

An initial egglay was performed at 16°C overnight using adult gravid worms. Egglay was achieved on RNAi plates and first generation progeny were maintained on those plates and at 16°C until gravid adult stage. Second generation egglay was performed at 16°C onto RNAi plates using first generation gravid adults to maximize RNAi efficiency (*hcf-1* RNAi generally has low efficiency and generally works best under these conditions and the RNAi bacteria seeding protocol described below). After the second generation egglay, plates were immediately moved to 25°C for remainder of experiment. Worms were scored a little less than 2 days later when they had reached L4/young adult stage. Approximately 50 worms were placed per slide for mounting and GFP localization in the intestines was scored using a fluorescence microscope. Note that for the experiments involving *skn-1(zu67);gcs-1::gfp* worms, these worms are *unc* rollers and

extremely sterile. The *unc* acts as a balancer for the *skn-1* mutation. Worms homozygous for *skn-1(zu67)* are sterile but phenotypically roller. In order to obtain *skn-1* mutants for the experiments, egglay was performed of about 12-15 *unc* roller gravid adults on 15 plates. I then scored only roller worms, which because of the balancer must be *skn-1* mutants. Despite performing such a large egglay, very few of the progeny were rollers, and so the N for this experiment was small. This strain is quite sterile, and so despite a large egglay, the majority of the eggs failed to hatch and those that did were disproportionally heterozygous for the *skn-1* mutation rather than homozygous. The reason for the lack of *skn-1* homozygotes is currently unknown, but the ratio of homozygotes to heterozygotes was far smaller than the expected 1:2.

Nuclear Localization Experiments -* tert-*Butyl hydroperoxide (t-BOOH)

Adult gravid worms were allowed to lay eggs on stock plates and the progeny were grown at 25°C until young adult/early gravid adult and then transferred onto stock plates with 2mM t-BOOH. 2mM t-BOOH was added to fresh stock plates the day before the worms were transferred. Worms were left on t-BOOH 2-3 hours before scoring nuclear SKN-1::GFP localization using a fluorescence microscope. Scoring was performed as indicated above under *hcf-1 RNAi*.

Nuclear Localization Experiments - Sodium Arsenite

Adult gravid worms were allowed to lay eggs on stock plates and the progeny were grown at 25°C until young adult/early gravid adult and then transferred onto stock plates with 10mM NaAs. Plates were prepared by adding 0.130g Sodium MetaArsenite in 2mL of M9, covering the tube with foil and keeping on rotator for ~10 minutes until dissolved. 110uL of this solution was added per 60mm stock plate. Worms were left on the plate for one hour before scoring. Scoring was performed as indicated above under *hcf-1 RNAi*.

Stock plate seeding protocol and strain maintenance

Stock plates were created by growing *E. coli* OP50 bacteria overnight at 37°C. OD was measured and culture was concentrated to OD 7.5 directly before adding to 60mm Nematode Growth Medium (NGM) plates.

Strains were kept on stock plates and chunked every 2-3 days (strains with normal fertility) and 5-6 days (strains with limited fertility – *rrf-3;hcf-1*) in order to prevent starvation for at least two weeks (or equivalently at least four generations) prior to being used in an experiment.

RNAi bacteria seeding protocol

HT115 competent cells were transformed with vectors expressing dsRNA and grown at 37°C in LB containing 100 ug/mL carbenicillin and 15 ug/mL until OD 0.8. IPTG was added to the OD 0.8 cultures until 4mM and cultures were induced on the bench for 4 hours. Culture was then concentrated to OD 7.5 and 35mm NGM plates were seeded with 150uL culture each and dried at room temperature. Before plates were used in an experiment, they were induced with 4mM IPTG and left to induce overnight.

Microarray data analysis

Microarray data for two sets of arrays were obtained from Gizem Rizki. The first set was comprised of two arrays to compare *hcf-1* mutant worms on the first channel and wild-type N2 worms on the second channel. The second set of 7 arrays compared wild-type worms (*skn-1(+)*) on the first channel with *skn-1* RNAi treated wild-type worms on the second channel. The data were log transformed and merged into a single dataset. Genes with a positive log(base 2) value in the *hcf-1* arrays are up-regulated in the absence of HCF-1 (e.g. upregulated in the *hcf-1* mutant worms when compared to wild-type), while genes with a positive value in the *skn-1* arrays are up-regulated in the presence of SKN-1 (e.g. upregulated in the wild-type worms when compared to wild-type treated with *skn-1* RNAi).

The Signficance Analysis of Microarrays (SAM) platform was used to identify genes expressed similarly in *hcf-1(-)* and *skn-1(+)*, as well as genes expressed differently between these two groups (Tusher *et al*. 2001). To find genes that are significantly similar, I performed one class analysis across all arrays, with a false discovery rate (FDR) = 0. The FDR is a statistical prediction of the number of genes called that are expected to be false positives. Since many genes were identified as significant even under this stringent FDR, an FDR of 0 was used. To find genes that are significantly different, I

performed two class unpaired analyses, again with FDR = 0. Finally, SAM was used to find genes significantly changed in just *hcf-1(-)* or just *skn-1(+)*, again with one class analysis using an FDR of 0. The entire microarray dataset was then filtered so that only genes flagged in at least one of the SAM runs were kept for subsequent analyses. Hierarchical clustering was performed on these remaining genes to obtain the heatmap shown in Fig. 5. Functional annotation clustering was performed by submitting the list of genes in each cluster shown in Fig. 5 to DAVID (Huang *et al*. 2009).

REFERENCES

An, J. H. and Blackwell, T. K. (2003). SKN-1 links *C. elegans* mesendodermal specification to a conserved oxidative stress response. *Genes & Dev,* **17**: 1882-1893.

Harman, D., 1956. Aging: a theory based on free radical and radiation chemistry. *J. Gerontol*. **11**: 298–300.

Honda, Y. and Honda, S. (1999). The daf-2 gene network for longevity regulates oxidative stress resistance and Mn-superoxide dismutase gene expression in Caenorhabditis elegans. *FASEB J*. **13**(11): 1385-93.

Huang DW, Sherman BT, Lempicki RA. (2009). Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nature Protoc*. 4(1):44-57.

Huang DW, Sherman BT, Lempicki RA. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*. 37(1):1-13.

Hyun, M., Lee, J., Lee, K., May, A., Bohr, V. A. and Ahn, B. (2007). Longevity and resistance to stress correlate with DNA repair capacity in *Caenorhabditis elegans*. *Nucl Acids Res*. **36**(4): 1380-1389.

Kenyon, C., Chang, J., Gensch, E., Rudner, A. and Tabtlang, R. (1993). A *C. elegans* mutant that lives twice as long as wild-type. *Nature,* **336**: 461-464.

Lee, S. S., Kennedy, S., Tolonen, A., and Ruvkun, G. (2003). DAF-16 target genes that control *C. elegans* lifespan and metabolism. *Science,* **300**: 644-647.

Lee, S., Horn, V., Julien, E., Liu, Y., Wysocka, J., Bowerman, B., Hengartner, M. O. and Herr, W. (2007). Epigenetic regulation of histone H3 serine 10 phosphorylation status by HCF-1 proteins in *C. elegans* and mammalian cells. *PLoS ONE*, **2**(11).

Li, J., Ebata, A., Dong, Y., Rizki, G., Iwata, T. and Lee, S. S. (2008). Caenorhabditis elegans HCF-1 functions in longevity maintenance as a DAF-16 regulator. *PLoS Biol* **6**(9): e233.

Liu, Y., M. O. Hengartner, and W. Herr. (1999). Selected elements of herpes simplex virus accessory factor HCF are highly conserved in Caenorhabditis elegans. *Mol. Cell. Biol.* **19**:909–915.

Ogg, S., Paradis, S., Gottlieb, S., Patterson, G.I, Lee, L., Tissenbaum, H.A., Ruvkun, G. (1997). *Nature* **389**(6654): 994-999.

Oliveira, R. P., Porter Abate, J., Dilks, K., Landis, J., Ashraf, J., Murphy, C. T. and Blackwell, T. K. (2009). Condition-adapted stress and longevity gene regulation by Caenorhabditis elegans SKN-1/Nrf. *Aging Cell* **8**(5): 524-41.

Rains, J. L. and Jain, S. K. (2011). Oxidative stress, insulin signaling, and diabetes. *Free Radic Biol Med*. **50**(5): 567-575.

Reuter, S., Gupta, S. C., Chaturvedi, M. M. and Aggarwal, B. B. (2010). Oxidative stress, inflammation, and cancer: how are they linked? *Free Radic Biol Med* **49**(11): 1603-16.

Rizki, G., Iwata, T. N., Li, J., Riedel, C. G., Picard, C. L., Jan, M., Murphy, C. T. and Lee, S. S. (2011). The evolutionarily conserved longevity determinants HCF-1 and SIR-2.1/SIRT1 collaborate to regulate DAF-16/FOXO. *PLoS Genetics*. **7**(9).

Tullet, J. M., Hertweck, M., An, J. H., Baker, J., Hwang, J. Y., Liu, S., Oliveira, R. P., Baumeister, R. and Blackwell, T. K. (2008). Direct inhibition of the longevity-promoting factor SKN-1 by insulin-like signaling in C. elegans. *Cell* **132**(6): 1025-38.

Tusher, V. G., Tibshirani, R. and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *PNAS* **98**(9): 5116-5121.

Wysocka, J., Reilly, P. T., Herr, W. (2001). Loss of HCF-1-chromatin association precedes temperature-induced growth arrest of tsBN67 cells. *Mol Cell Biol.* **21**(11): 3820-9.

FIGURES AND TABLES

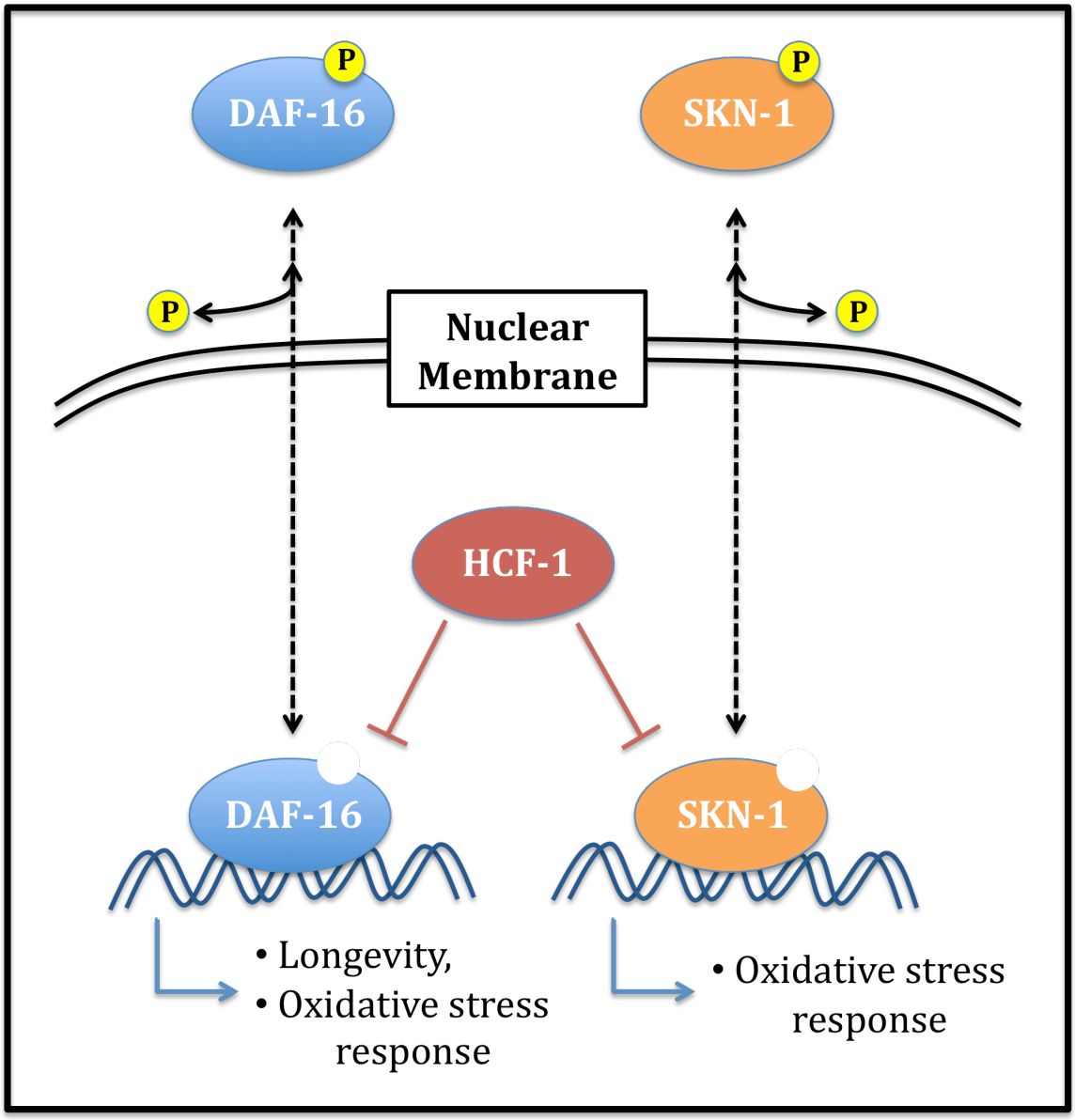


Fig. 1: Our results show that HCF-1 acts as a suppressor of both DAF-16 and SKN-1-mediated oxidative stress response pathways, as well as DAF-16 mediated longevity.

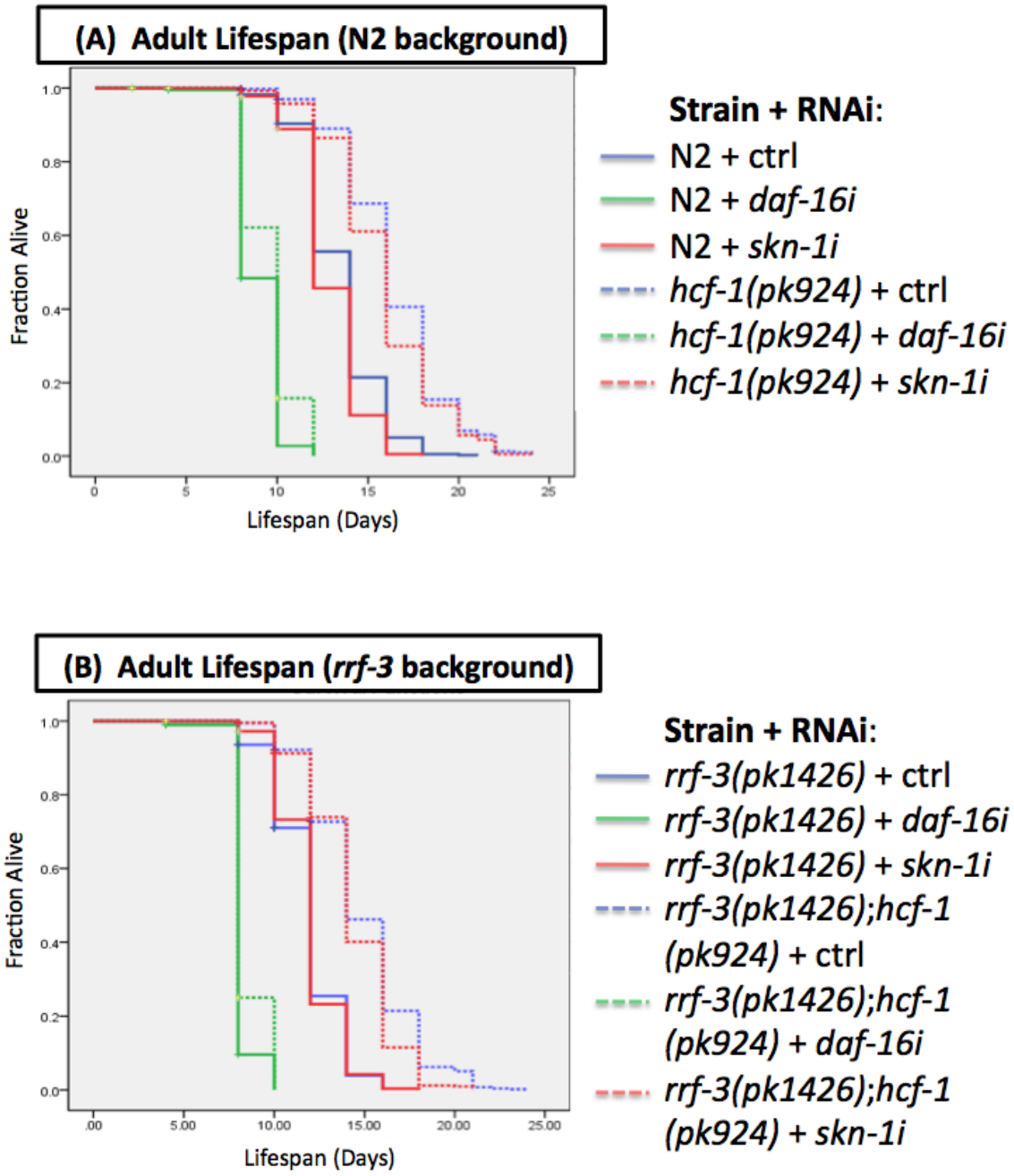


Fig. 2: SKN-1 knockdown causes comparable small lifespan decreases in both wild-type (N2) and *hcf-1*(-) backgrounds. The same is true when performed in the *rrf-3* RNAi sensitive background. Worms were treated with the indicated gene targeting or control RNAi starting at egglay. **(A)** Data from two experiments including *daf-16* RNAi and four experiments total were pooled. **(B)** Data from one experiment including *daf-16* RNAi and three experiments total were pooled. For summary statistics please refer to Table 1.

Table 1A: Lifespan data comparing average longevity of wild-type or *hcf-1(-)* mutant worms subjected to either control, *daf-16* or *skn-1* RNAi

Strain + RNAi	Mean Survival + SEM (Days)	Total N	*p*-value vs. N2 + ctrl	*p*-value vs. *hcf-1(pk924)* + ctrl	% effect on N2 + ctrl	% effect by *hcf-1(pk924)* vs. corresponding N2 + RNAi
N2 + ctrl	13.417 ± 0.11	392		<0.001		
N2 + *daf-16*	9.001 ± 0.08	184	<0.001	<0.001	-33	
N2 + *skn-1*	12.880 ± 0.09	394	<0.001	<0.001	-4	
hcf-1(pk924) + ctrl	16.358 ± 0.15	388	<0.001		22	22
hcf-1(pk924) + *daf-16*	9.558 ± 0.14	189	<0.001	<0.001	-29	6
hcf-1(pk924) + *skn-1*	15.839 ± 0.08	398	<0.001	0.014	18	23

Data shown are pooled from four independent experiments, two of which included *daf-16* RNAi and two of which did not. Adult gravid worms of the indicated strain were allowed to lay eggs overnight at 16°C onto prepared and induced RNAi plates targeting the indicated genes. Following egglay, worms were transferred to 25°C for the remainder of the experiment. Worms were transferred onto fresh RNAi plates containing FUDR at the young adult/early gravid adult stage ('Day 0') and were subsequently transferred onto new RNAi + FUDR plates at Day 2, 4 and 8 of adulthood. A survival plot of these data shown in Fig. 2A.

Table 1B: Lifespan data comparing average longevity of wild-type or *hcf-1(-)* mutant worms in RNAi sensitive *rrf-3* background, subjected to either control, *daf-16* or *skn-1* RNAi

Strain + RNAi	Mean Survival + SEM (Days)	Total N	*p*-value vs. N2 + ctrl	*p*-value vs. *hcf-1(pk924)* + ctrl	% effect on N2 + ctrl	% effect by *hcf-1(pk924)* vs. corresponding N2 + RNAi
rrf-3(pk1426) + ctrl	11.879 ± 0.11	296		<0.001		
rrf-3(pk1426) + *daf-16*	8.151 ± 0.07	98	<0.001	<0.001	-31	
rrf-3(pk1426) + *skn-1*	11.966 ± 0.10	288	0.810	<0.001	1	
rrf-3(pk1426);hcf-1(pk924) + ctrl	14.827 ± 0.12	540	<0.001		25	25
rrf-3(pk1426);hcf-1(pk924) + *daf-16*	8.500 ± 0.09	101	<0.001	0.001	-28	4
rrf-3(pk1426);hcf-1(pk924) + *skn-1*	14.359 ± 0.11	435	<0.001	<0.001	21	20

Data shown are pooled from three independent experiments, one of which included *daf-16* RNAi and two of which did not. In addition, data from three independent *rrf-3(pk1426);hcf-1(pk924)* double mutant lines have been combined. Adult gravid worms of the indicated strain were allowed to lay eggs overnight at 16°C onto prepared and induced RNAi plates targeting the indicated genes. Following egglay, worms were transferred to 25°C for the remainder of the experiment. Worms were transferred onto fresh RNAi plates containing FUDR at the young adult/early gravid adult stage ('Day 0') and were subsequently transferred onto new RNAi + FUDR plates at Day 2, 4 and 8 of adulthood. A survival plot of these data shown in Fig. 2B

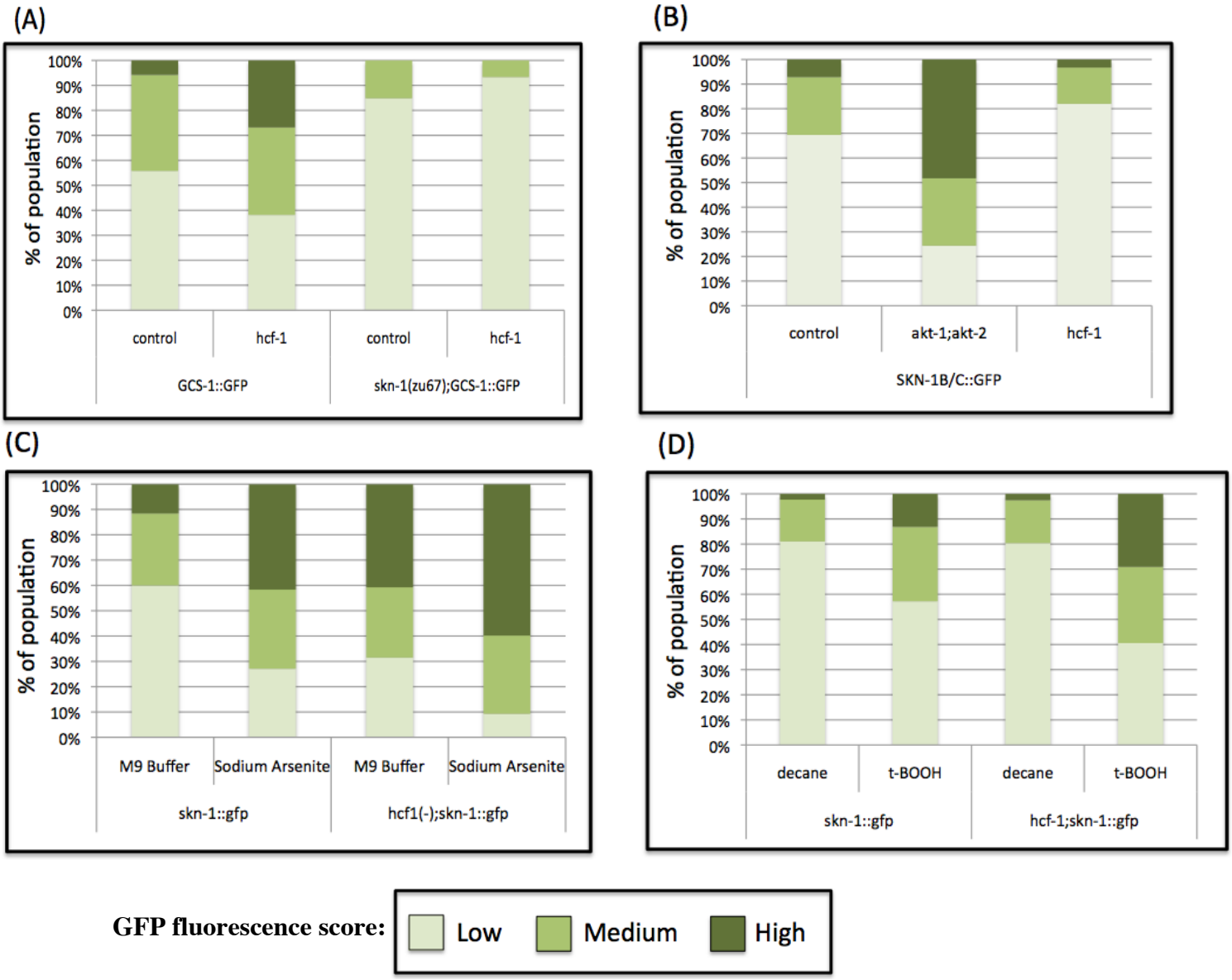


Fig. 3: **(A-B)** SKN-1B/C::GFP, GCS-1::GFP and *skn-1(zu67)*;GCS-1::GFP worms were treated with either control, *akt-1* and *akt-2* or *hcf-1* RNAi at egglay and were grown until L4. L4 worms were scored under a fluorescence scope with a GFP filter. Worms were scored as having 'high' GFP localization if a high signal of GFP was present throughout the intestine (GCS-1::GFP) or intestinal nuclei (SKN-1B/C::GFP). Worms scored as 'medium' had GFP signal anteriorly, posteriorly, or both, but no signal in the middle of the intestine. Worms scored as 'low' had no visible GFP in the intestine. See Fig. 4 for illustration of scoring criteria. Data for **(A)** represent two independent experiments, while data for **(B)** represent a single experiment. **(C)** Adult gravid worms were allowed to lay eggs on stock OP50 plates and progeny were grown until young adult (YA)/early gravid adult (GA) stage and then transferred onto OP50 plates with 10mM NaAs. Worms were scored after 1 hour on NaAs. Results represent a single experiment. **(D)** Egglay also on OP50 and worms were grown until YA/GA before transferring to OP50 plates with 2mM t-BOOH. Worms were scored after 2-3 hours on t-BOOH. Results shown are pooled from two similar experiments. **(A-D)** All experiments shown in this figure were scored blind, meaning the experimenter did not know which strains and which conditions were being evaluated while scoring. See Tables 2 and 3 for tabulated results and p-values.

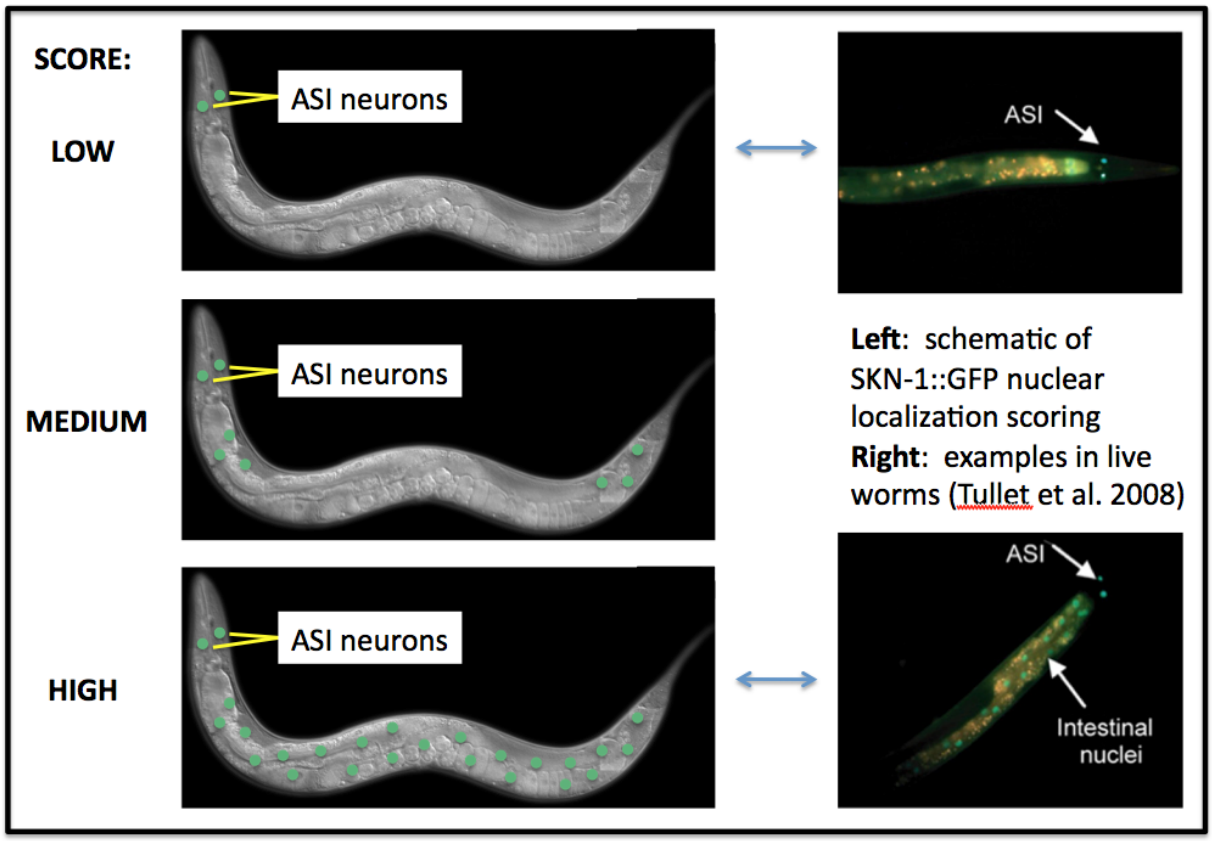


Image credits: (LEFT) Ian D. Chin-Sang (Queen`s University, Kingston, ON, Canada); (RIGHT) Tullet et al. 2008

Fig. 4: SKN-1::GFP nuclear localization scoring scheme used in fluorescence microscopy assays. SKN-1 nuclear localization was scored as low if no GFP signal was visible in the intestine (note that SKN-1 is constitutively expressed in the ASI neurons). A score of medium indicates that nuclear SKN-1 was visible anteriorly, posteriorly, or both, but with no signal in the middle of the intestine (middle panel). A score of high was used to indicate strong nuclear SKN-1 signal throughout the intestine (bottom panel). For experiments using GCS-1::GFP, the score was similarly based on the amount of signal visible in the intestine (not shown). GCS-1 was considered highly induced if a strong GFP signal was visible throughout the intestine. Medium induction indicates a GFP signal anteriorly, posteriorly, or both, but not in the middle of the intestine, and Low induction indicates little to no GFP visible in the intestine. Note that the yellowish fluorescence in the images to the right is due to autofluorescence in the gut of worms. The green signal represents the GFP signal from SKN-1::GFP.

Table 2: HCF-1 does not affect SKN-1 nuclear localization under basal conditions, but does affect the nuclear localization of known SKN-1 target gene GCS-1. SKN-1 is required for GCS-1 localization into the intestine.

Strain	RNAi	GFP Fluorescence			
		Low	Medium	High	Total
skn-1b/c::gfp	**control**	136	46	14	196
	akt-1;akt-2	49	55	97	201
	hcf-1	150	27	6	183
gcs-1::gfp	**control**	58	40	6	104
	hcf-1	37	34	26	97
skn-1(zu67); gcs-1::gfp	**control**	28	5	0	33
	hcf-1	14	1	0	15

Data shown represent two pooled experiments for the SKN-1B/C::GFP assays, and a single experiment of the GCS-1::GFP and *skn-1(zu67)*; GCS-1::GFP assays. Worms were grown on the indicated RNAi at 25°C until L4 before scoring as indicated in Fig. 4 and the materials and methods section. No significant difference was found for SKN-1 nuclear localization in *hcf-1* background relative to control. *hcf-1* caused the induction of SKN-1 target GCS-1::GFP, which was fully suppressed in *skn-1* mutant background.

Table 3: SKN-1 is further induced into the intestine as a result of *t*-BOOH or NaAs exposure in *hcf-1* background compared to wild-type.

Strain	Condition	Degree of Nuclear SKN-1::GFP				*p*-values	
		% Low	% Med	% High	Total N	relative to control	relative to wild-type
skn-1::gfp	**decane**	81.0	16.7	2.3	174		
	t-BOOH	57.2	29.7	13.1	236	<0.001	
hcf-1(-);skn-1::gfp	**decane**	80.4	17.1	2.5	158		0.7740
	t-BOOH	40.6	30.4	29.1	313	<0.001	<0.001
Strain	**Condition**	**Degree of Nuclear SKN-1::GFP**				***p*-values**	
		% Low	**% Med**	**% High**	**Total N**	**relative to control**	**relative to wild-type**
skn-1::gfp	**M9 Buffer**	60.0	28.4	11.6	95		
	NaAs	27.1	31.3	41.7	96	<0.001	
hcf-1(-);skn-1::gfp	**M9 Buffer**	31.5	27.8	40.7	108		<0.001
	NaAs	9.3	30.9	59.8	97	<0.001	<0.001

Data shown for *t*-BOOH represent the pooled results of two separate experiments. In addition, *t*-BOOH experiments were performed with two separate *hcf-1(pk924);skn-1::gfp* strains, which behaved very similarly and are pooled in these results. Data shown for NaAs represents a single experiment. GFP fluorescence was scored as indicated in Fig. 4 and the materials and methods section.

Table 4: Gene Ontology clustering results for genes similarly changed in *hcf-1(-)* and *skn-1(+)*, as well as genes oppositely affected in the two groups, and genes affected only by *hcf-1(-)*.

Genes similar in *hcf-1(-)* and *skn-1(+)* (cluster C in Fig. 4)				
GO Term	**General Process**	**Enrichment Score**	**p-value**	**Benjamini**
Glutathione S-transferase	Stress Response	8.7	5.2E-15	1.3E-12
Peptidases	Proteolysis	3.12	3.0E-5	0.0016
Serine Hydrolase	Proteolysis	2.77	3.0E-5	0.0016
Aspartic-type peptidase	Proteolysis	2.32	8.4E-4	0.029
Lysozyme Activity	Metabolism	2.30	2.8E-4	0.035
Genes opposite in *hcf-1(-)* and *skn-1(+)* (clusters A and E in Fig. 4)				
GO Term	**General Process**	**Enrichment Score**	**p-value**	**Benjamini**
Multicellular organismal aging	Aging	2.33	0.0047	0.20
Amino acid catabolism	Metabolism	2.3	3.3E-4	0.11
Vitamin B6 Binding	Metabolism	2.23	7.1E-4	0.096
Peptidases	Proteolysis	2.13	7.9E-4	0.072
Coenzyme metabolism	Metabolism	1.94	0.0044	0.27
Genes affected by *hcf-1(-)* only (clusters B and D in Fig. 4)				
GO Term	**General Process**	**Enrichment Score**	**p-value**	**Benjamini**
Nematode Cuticle Collagen	Structural Component	3.88	5.2E-6	0.003
Multicellular organismal aging	Aging	3.09	8.1E-4	0.35
Cytoskeleton	Structural Component	2.96	1.4E-5	0.0039
Lipid metabolism	Metabolism	1.87	0.0059	0.29
Drug metabolism	Metabolism	1.82	2.9E-5	0.0018

The list of genes from each cluster indicated on Fig. 5 was submitted to DAVID for Functional Annotation clustering. The top 5 clusters, based on enrichment scores, are reported for the genes similarly changed in *hcf-1(-)* and *skn-1(+)*, the genes oppositely changed in *hcf-1(-)* and *skn-1(+)*, and genes affected only by *hcf-1*. For each cluster, the most significant p-value is reported, along with the Benjamini-Hochberg corrected p-value ("Benjamini").

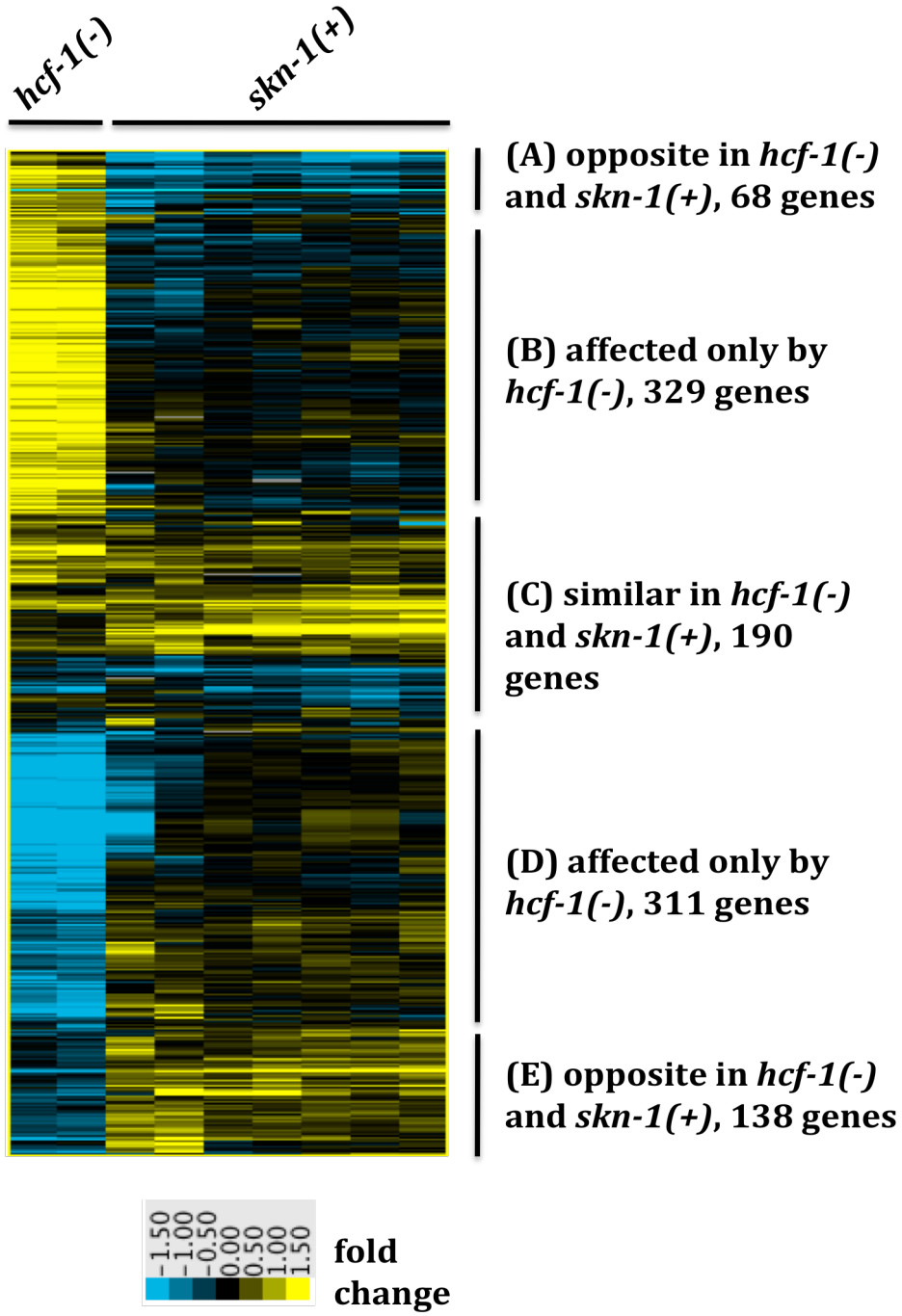


Fig. 5: Microarray data for two arrays comparing *hcf-1* mutants to wild-type worms, and seven arrays comparing wild-type worms treated with control RNAi to *skn-1* RNAi. The *hcf-1(-)* arrays are positive (yellow) for genes whose mRNA levels increased in *hcf-1* background relative to wild-type, while the *skn-1(+)* arrays are positive (yellow) for genes whose mRNA levels were higher in wild-type worms on control RNAi than on *skn-1* RNAi.

Chapter 2: Determination of enrichment regions for H3K27me3 and other low-signal, high-noise ChIP-seq data

ABSTRACT

Chromatin modifications are a major mechanism through which cells regulate gene expression. To study these mechanisms, researchers are increasingly making use of a technique that combines chromatin immunoprecipitation (ChIP) with recent advances in massively parallel DNA sequencing, in a process called ChIP-seq. This approach allows for genome-wide profiling of any DNA-binding protein, including histones and transcription factors. To analyze these data, researchers have developed statistical tools that can identify genomic regions enriched for the DNA-binding protein of interest. In particular, transcription factors and some other proteins produce data profiles with sharp peaks. The many programs whose purpose is to identify these sharp peaks are referred to as "peak-callers". Some histone marks, however, bind across broad regions of the genome and produce diffuse and high-noise data profiles that are often difficult to analyze. In this paper, I investigate the ability of peak-callers to analyze ChIP-seq data from H3K27me3, a histone mark known to produce broad regions of enrichment. Further, I present an alternative method that is both faster and simpler than the than the other peak callers investigated. This alternative method focuses on genomic features of interest, such as genes or promoters, directly. It offers promise in identifying enriched genomic features when a histone mark, like H3K27me3, generates very diffuse and noisy ChIP-seq data patterns.

INTRODUCTION

Chromatin modifications are used by cells to regulate gene expression by controlling protein access to the promoter regions and other DNA in the genome. Different modifications can lead to different transcriptional outcomes by causing conformational changes that either allow or prevent proteins to access the DNA, or by altering the recruitment of effector protein complexes (Park 2009). As it has become increasingly clear that chromatin modifications play a major role in the already complex mechanisms surrounding gene expression, there has been an explosion in the use of ***ch****romatin* ***i****mmuno****p****recipitation* (ChIP) followed by deep sequencing (ChIP-seq) to study these mechanisms. ChIP-seq can be used to reveal where a protein of interest binds to its target DNA, including finding transcription factor binding sites and RNA polymerase targets. While the historical use of ChIP-seq was primarily for the discovery of novel transcription factor binding sites, this type of experiment is increasingly used to study the role of histone modifications in diverse situations (Park 2009).

In a ChIP-seq experiment, DNA is reversibly bound to its associated proteins to form a complex in a process called cross-linking. The DNA-protein complex is then sheared, and an antibody is used to enrich for a protein of interest. DNA fragments associated with the protein of interest will also become enriched in the solution due to the crosslinking. The DNA-protein bonds then are removed and the resulting solution is purified to produce a solution of DNA, which can be sequenced. After mapping all the fragments ('reads') obtained from sequencing to a reference genome, the experiment should produce a significant increase in the number of fragments in locations where the protein of interest can be found bound to its target DNA.

The ability to generate data and the volume of data generated in ChIP-seq experiments have, in some cases, outstripped the development of statistical theory and methods for their analysis. The sheer volume and variability of some types of ChIP-seq data pose a significant challenge for data analysis, as researchers attempt to robustly identify the specific regions that are significantly enriched for the proteins of interest. Generally, this is done by comparing the number of sequenced reads that map to a region in the ChIP sample to the number that map to the same region from a background or control sample (Wilbanks and Facciotti 2010, Park 2009, Rozowsky et al. 2009).

Enriched regions are usually referred to as 'peaks' and the programs that find them are correspondingly called 'peak-callers.'

In the case of proteins such as transcription factors, which bind strongly and precisely in small (usually < 20 bp) regions of DNA, ChIP-seq produces small regions of very high enrichment relative to the control. These are usually referred to as 'sharp peaks,' due to their appearance when viewing the data in a genome browser such as IGV (see Fig. 1) (Robinson et al. 2011). This makes these regions relatively easy to identify, both visually and computationally. Because ChIP-seq was originally used primarily for these kinds of analyses, all published peak-callers can usually identify such peaks. A study comparing various algorithms for peak-calling in the context of transcription factor ChIP-seq data failed to find compelling evidence that a particular algorithm is better than another, concluding that all are able to identify sharp peaks within an acceptable margin of error (Wilbanks and Facciotti 2010). In other words, most if not all published peak-callers should work if analyzing data that is composed primarily of sharp peaks. Since the conclusions of Wilbanks and Facciotti (2010) suggest that all peak callers are equivalent with respect to their ability to call peaks, the authors suggest that researchers choose peak callers based on ease of use. Some peak callers are much easier to use than others, due to better interfaces or increased stability, while others are quite difficult to interact with and will crash regularly. It is therefore advised to use a more stable, well-supported peak-calling program when looking for sharp peaks in ChIP-seq data. There is, however, no doubt that the problem of finding sharp peaks in ChIP-seq data has been abundantly and well addressed in the literature, and that anyone performing this type of data analysis will be able to find a suitable peak-caller among those already published.

Although any available peak caller should be able to identify sharp peaks, the same is not true when assaying for proteins that bind across broad – and less well-defined - regions of the genome. The histone mark that is the primary focus of this paper, a tri-methylation of the 27th lysine of the H3 histone protein subunit (H3K27me3), is a broad suppressive mark whose ChIP-seq data have been somewhat notoriously difficult to analyze. I have obtained H3K27me3 data from ChIP-seq experiments performed in *Caenorhabditis elegans* (Zoey Ni, unpublished data). These data exhibit both **weak signal** and **high noise**, in addition to being broadly distributed with no sharp, easily

identifiable peaks. These characteristics make it extremely difficult to assess, or even define the notion of, enrichment for these data. Most peak-callers, particularly those designed for sharp peaks, such as PeakSeq (Rozowsky et al. 2008) or MACS (Zhang et al. 2008), are unable to identify broad regions of enrichment because the assumptions in their underlying statistical models simply do not hold in these types of data. In particular, the algorithms tend to look for strong local enrichment, and therefore fail to identify the more diffuse signals of H3K27me3 modification (Rozowsky et al. 2008, Zhang et al. 2008).

More recently, some programs have been developed to identify broader regions of enrichment (Xu et al. 2008, Zang et al. 2009, Rashid et al. 2011). Given this, it is worth exploring how these new peak-callers work with the H3K27me3 data. Here I present the results of preliminary analysis of the H3K27me3 data using ChIPDiff (Xu et al. 2008), SICER (Zang et al. 2009) and ZINBA (Rashid et al. 2011), all peak callers optimized to identify enriched regions in ChIP-seq data with broad marks. I also present a simple alternative method that I believe shows promise for analyzing these data and other similar ChIP-seq data.

PRELIMINARY PEAK-CALLING RESULTS

I assessed the performance of three published peak-callers, designed specifically to identify broad regions of enrichment, on the H3K27me3 data. First I give a broad overview of the different algorithms used by these programs, and then I present the results of the peak calling.

ChIPDiff (Xu et al. 2008)

This algorithm uses a Hidden Markov Model (HMM) to parse the genome into regions of enrichment and non-enrichment. As is the case for most peak callers, ChIPDiff uses a two-pass approach. In the first pass, the genome is divided into non-overlapping consecutive bins of default length 1000bp. The number of reads whose centers (assuming a fragment length of 200bp) fall in each bin is counted, and putative regions of enrichment are roughly identified using a minimum fold change ('putative bins'). Putative regions separated by less than *1000bp* are then grouped together into

broader regions. In the second pass, the putative regions are more rigidly evaluated under the HMM statistical model to identify sub-regions of significantly different enrichment levels between sample and control. The program returns both regions that are significantly enriched for the sample relative to the control, and regions significantly enriched for the control relative to the sample.

***SICER* (Zang et al. 2009)**

The SICER algorithm uses a similar binning (here called 'windows' – default 200bp) and two-pass strategy as ChIPDiff. In the first pass, the program identifies 'eligible' windows as windows with a read count higher than a specific threshold. Groups of 'eligible' windows separated by less than *g* 'ineligible' windows are considered a cluster, which is the unit of all subsequent statistical analysis. In the second pass, the 'scores' of each cluster are computed as a function of the read counts, and all clusters with a score above a particular threshold, determined by the desired FDR and calculated by determining the probability of observing a cluster with that score, are reported as enriched.

***ZINBA* (Rashid et al. 2011)**

ZINBA also uses a two-pass strategy. In the first pass, the genome is parsed into non-overlapping windows with a default length of 250bp. The program then counts the number of reads in each one and assigns a 'score' to each window. The second pass, as in most algorithms, applies the statistical model (here a novel mixture regression model) to classify each window as background, enriched, or zero-inflated. The zero-inflated condition represents windows where the actual number of reads mapped is inferior to the number we would expect given the coverage, and may represent areas containing a high proportion of unmappable bases, or more generally a lack of sequencing depth. Also notable is that ZINBA does not assume a single statistical model for the background. Instead, the algorithm uses a common criterion, the Bayesian Information Criterion (BIC), which is a goodness of fit measure. The BIC is used to find the most appropriate model for the background from a finite set of possible models. This would theoretically allow ZINBA to better capture the background distributions of ChIP-seq data, which are

often quite variable. A part of the statistical model used by ZINBA is designed specifically for identifying enriched regions in low signal-to-noise data.

Results

The regions considered enriched by each algorithm discussed are shown in a small, representative portion of chromosome I in Fig. 2 and supplemental Fig. 1. It is immediately clear by visual inspection that both ZINBA and ChIPDiff perform poorly at identifying the broad enriched domains in these data. Both the ChIPDiff and ZINBA enriched regions are small and sparse, and visually are not consistent with the data. For ChIPDiff, this is likely because the first pass eliminates the majority of the bins in the genome. There is simply not enough total enrichment in most bins for them to be considered putative using the simple cutoff in ChIPDiff. Therefore ChIPDiff appears to be inappropriate for data with low signal. The ZINBA results were more surprising given the adjustments the algorithm is supposed to make in response to low signal, but the program was quite unstable and crashed regularly, so the results are difficult to trust. SICER results are shown using a window size of 200bp and 1000bp. The former has higher resolution but results in a more 'staccato' pattern of peaks, while the latter results in larger – and potentially too large - enriched regions.

While visual inspection confirms that both versions of SICER perform better than either ZINBA or ChIPDiff, whether the SICER results are "right," and which of the two SICER settings – 200bp or 1000bp windows – is best, is difficult to determine. When data consist of sharp peaks, which are relatively clear and easy to identify, evaluating whether a peak caller was able to identify those peaks is usually simple to do by eye. However, there is no established way of determining whether a particular peak-caller actually 'works' with data that are less clean-cut. In particular, it is difficult to determine what the boundaries for a particular enriched region should be (Fig. 3). In Fig. 3 we see that in (A), a small region of negative enrichment fails to interrupt an otherwise positive region of enrichment. However, in (B) we see that a similar small negative region causes the same peak caller, SICER, to separate the positive region into two distinct regions in order to exclude the negative region. This means that the SICER results are not consistent even within a single run of the algorithm, treating what appear to the naked eye

to be similar regions differently. The exact meaning of these differences is unclear. What is the significance of excluding the negative region in (B) while keeping the negative region in (A)? How should this result be interpreted? There is very little literature addressing the evaluation of peak callers for this type of data, and so very little guidance available to the researcher. Perhaps more to the point, papers that are not technical or methods papers, but use a peak-caller as a part of their data analysis pipeline, give little or no justification as to why they chose a particular peak-caller beyond citing it (Akkers et al. 2009, Rada-Iglesias et al. 2011, and Maruyama et al. 2011, for example).

Finally, analyzing the results of any of these peak-callers in the context of broad peaks is non-trivial. The purpose of most research using ChIP-seq is not to find the enriched regions themselves, but to find the significance of those regions, to "interpret" the data. If a researcher is looking for genes (or any other genomic feature) that are enriched for the broad marker in question, a common part of the downstream analysis of peak-calling data (Akkers et al. 2009, Young et al. 2011, among others), they must determine how they will translate the peak-calling results into meaningful results that can be analyzed. Should every gene (or other genomic feature) with more than a certain degree of overlap with an enriched region be considered enriched? Or should only genes that are contained entirely within an enriched region be considered for further analyses? The goals and approach used for the downstream, post-peak-calling data analysis may provide some guidance as to what may be the most appropriate method to identify enriched regions. With this in mind, in the remainder of this paper, I discuss an alternative methodology that may prove useful when the data analysis pipeline aims to identify enrichment across specific genomic features, and may perhaps also be used for comparing ChIP-seq data across experiments.

A DIFFERENT METHOD

In an attempt to address the questions raised by the peak-calling results just presented, I created a simple but different way to identify enriched regions and tested it on the H3K27me3 data.

Briefly, in this algorithm, the data are binned with respect to the genomic feature(s) of interest, and then each of these bins are evaluated separately for enrichment

in the sample relative to the control using a simple t-test. For example, if the genomic feature of interest is exons, each exon becomes a bin. The algorithm counts the number of reads in each exon from the sample, and the number from the control, and then uses these values to perform a simple t-test to determine whether there are more reads in the bin from the sample then from the control. The binning using a genomic feature of interest just described was accomplished two different ways, and both are explored here. The algorithms are explained in further detail in the methods section.

Binning using per-base-pair counts

In the first method, the original data, which contained a chromosome identifier and the start position and direction of each tag, were manipulated in order to obtain per-base-pair tag counts. Each base pair was assigned a value corresponding to the number of reads overlapping at that location. Ideally, we would calculate this using the start and end position of a sequenced read, and increment the count for all base pairs between the start and end position by one (as shown in Fig. 5). This would be repeated for all the reads to produce the density data. Since sequencing data does not actually contain the end position of a tag, this method cannot be used directly. However, size-selection is used before sending ChIP-seq DNA for sequencing, in order to remove all fragments outside a specific size range. In the case of the H3K27me3 data used here, fragments were size-selected to be between 150-300bp, with a median length of 200bp due to an excess of small fragments created by this process. This is standard for ChIP-seq experiments (Park 2009). Given this, I created a probabilistic distribution representing the fragment length as follows: for the first 150bps after a fragment's start position, the fragment is guaranteed to be present and occurs with probability 1. Furthermore, the fragment cannot be longer than 300bps so for >300bps after a fragment's start position, the probability of the fragment being present is zero. For all base pairs in between (150-300bps), I used a cumulative beta distribution (because such distributions have a finite support) and adjusted the parameters so that the median value would fall at 200bps (see methods). In other words, the probability of the fragment still being present at 200bps after the start position is approximately 0.5. An illustration of the probability density is given in Fig. 4(A), and the resulting distribution of tag lengths under this model is shown

in Fig. 4(B). In order to calculate the density, a vector of length 300 was created with the values shown in Fig. 4(A) – in other words, the first 150 values of the vector were 1, and the remainder were numbers between 0 and 1 as defined by the beta distribution. Then for each tag, the vector was placed at the start position of the tag and in the correct direction, and the counts in the vector were added to each base pair accordingly. This was repeated for all tags. The result is an estimate, for each base pair, of the (expected) number of fragments that overlapped with that base pair.

After expanding the data to a single base pair resolution, they were collapsed according to a set of genome annotations. The annotations used in this paper were simply the gene annotations of the ce6 *C. elegans* genome build, but the same method could be applied to any set (or sets) of genome annotations, including promoters, exons and introns. For each gene, the mean and standard deviation of the per-base-pair densities were calculated over all base pairs contained within the gene. The result was a set of genes with the mean and standard deviation of the per-base-pair densities that could be used to perform the required t-tests.

Binning using an intermediate binning step

The alternative way of binning according to gene annotations explored here is binning by way of an intermediate binning step. In this method the genome was parsed into consecutive non-overlapping bins of length L, and I counted the number of reads from the sample and control whose centers (assuming a median fragment length of 200bp) fell in each bin. This method is similar to the binning steps used by SICER, ChIPDiff and ZINBA. After this binning step, I determined which bins were in each gene by identifying the bin that the start and end positions of each gene fell into. This resulted in a range of bins that roughly described the location of each gene. I then calculated the average and standard deviation of the number of tag counts across all bins in each gene so I could perform the t-tests. Similarly to the per-base-pair binning method, this method can easily be used for any genome annotation information, but is demonstrated in the present paper using the ce6 gene annotations. Also, in this method the initial bin size can be varied, and was tested using both L = 200 and L = 1000 size bins (default bin sizes used in several peak caller programs).

Significance testing

Once I obtained the mean and standard variation of sample and control tag counts using either of the two methods above, I performed a simple t-test of difference of means in the sample and the control for each gene. For data obtained using the *per-base-pair counts* method of binning, the length of the gene (in base pairs) was used as the sample size N. For data obtained using the intermediate binning method, the number of bins in the gene was used for N (and genes with N = 1 were excluded as no standard deviation could be calculated). An unpaired, one-tailed t-test was used to test whether the number of reads in the sample was greater than in the control. Because I could not assume equal variances in the data, due in part to not knowing the expected background distribution, the test used Satterthwaite's formula to calculate the degrees of freedom appropriate for the test under the assumption of unequal variances (Satterthwaite 1946). Genes significant at the 0.005 level for the per-base-pair method (Fig. 6) and the 0.01 level for the intermediate binning method were identified as enriched in the sample, and exported for viewing in IGV.

Results

Visual inspection of the results of the per-base-pair binning method reveal that it shows some promise for identifying genomic features that are enriched for a particular marker using ChIP-seq data. The genes identified as being enriched clearly occur in areas of enrichment (corresponding to positive log (base 2) values calculated as described below; plotted in red in IGV in the first track in Fig. 6). In addition, the results compare well to those of SICER, and even occasionally appear superior (Fig. 7). This simple method is able to identify enriched genes in small, enriched areas that are otherwise too small for SICER to find (Fig. 7 (A) and (D)). In addition, the simple test of difference in means finds an almost identical region in 7(B), but has the advantage of leaving out a region in (C) which is potentially best left out because there are no genes contained wholly within that region. Again, although it is possible to interpret the SICER results in order to identify enriched genes, this process is non-trivial – the region identified in (C) overlaps with three genes, but the new method finds that none of these genes are significantly enriched. For researchers looking for enriched genes, this one-step process

to identify genes directly seems a much more revealing result than a region such as that defined by SICER.

The results of the new method using the intermediary binning method were similar to the per-base-pair binning results, though more conservative, particularly as bin size increases (Fig. 8). Notably, the intermediary binning method fails to identify small enriched genes when compared to the per-base-pair method (supplementary Fig. 2). However, the intermediary binning method is more conservative than the per-base-pair method, the latter of which calls some genes that do not visually appear enriched (Fig. 9). In summary, the per-base-pair method is very sensitive for small enriched genes, but may pick up a fair number of false positives in the case of large genes, as the large sample sizes (number of base pairs in the gene) can make small differences nonetheless statistically significant. The intermediate binning method is more conservative but fails to identify small enriched genes.

Finally, previous studies have shown that H3K27me3 preferentially marks inhibited genes (Young *et al*. 2011). Therefore, we expect a disproportionate number of genes enriched for H3K27me3 to have low mRNA levels. This was tested using RNA-seq data obtained from worms at the same time-point as the ChIP-seq data examined here (Mintie Pu, unpublished data). The RNA-seq data were used to place genes into one of three groups based on their expression levels – low, medium, or high (see methods). I then determined the degree of overlap between genes called by my method using intermediate binning with bin size 200, and genes with "low" or "high" expression levels. I performed the same analysis with the SICER results. In this case, I considered a gene "called" by SICER if at least 60% of the gene was contained in regions called by SICER. I reasoned that since the SICER results do not greatly overlap with my results (fig 10A), but visually in IGV appear similar, comparing the SICER results and my method's results using the RNA-seq data could be a meaningful way to determine which performs better. The results show that my method calls proportionally more genes with low expression levels than SICER, suggesting that my method more accurately captures genes enriched for H3K27me3 than does SICER (fig. 10B and 10C, table 1).

DISCUSSION

Results of preliminary analyses of the H3K27me3 data using ChIPDiff (Xu et al. 2008), SICER (Zang et al. 2009) and ZINBA (Rashid et al. 2011), all peak callers optimized to identify enriched regions in ChIP-seq data with broad marks, were found to be either poor, and/or difficult to interpret. Given these limitations, I developed and explored a simple alternative method for analyzing these data and other similar ChIP-seq data.

The new method of identifying enriched genomic features in ChIP-seq data is a preliminary effort, and should be further refined. In particular, there are assumptions about independence of observations that underlie the statistical tests that do not hold, particularly at the base-pair level. Because a single read can be anywhere from 150-300bp, adjacent base pairs will have very similar and highly correlated read counts. This correlation causes the standard deviation used in the t-test to be underestimated, or equivalently the N to be overestimated, given that the N base pairs in the gene do not constitute N independent observations. This increases the power of the statistical test, to the point where a large gene (with many thousands of base pairs for example) with an average tag count that is slightly higher in the sample than in the control will be considered enriched, despite the fact that visually that gene does not appear enriched. In fact, because of this large *N* issue, the per-base-pair binning method yields an almost binary distribution of genes – genes that are enriched with a p-value of zero, and those that are not with a p-value of one.

The intermediate binning method was conceived as a way to mitigate the effects of the correlations among the data points on the statistical test. This method, when $L = 200$, reduces the *N* to a reasonable approximation of the number of independent regions present in the gene, based on the fact that nucleosomes are known to be about 200bp in length (Gottesfeld and Melton, 1978). In a test with $L = 200$, *N* thus represents the approximate number of distinct nucleosomes present in the gene, and it is likely appropriate to consider these regions independent for our purposes. It is possible that other bin sizes could be meaningful, so the method was designed to work with any bin size. However, a bin size of less than 200bp is likely meaningless from a biological perspective, and very small bin sizes could create strange artifacts. Overall, this binning

method has the advantage of resulting in more conservative calling of enriched genes as seen in Figs. 8 and 9, and supplementary Figure 2. However, because small genes can contain very few bins, especially if a large *L* is used, this method loses the ability to detect enrichment in very small genes. The base-pair method, in contrast, is particularly adept at finding small enriched genes. It might therefore be advisable to create a hybrid-binning method, which uses the per-base-pair binning method for small genes (for example, genes smaller than 3 bins), and intermediate binning for all other genes.

The idea of simply comparing means for well-defined genomic features is particularly suited to deal with data where "enrichment" is ill defined, or very minimal. A typical peak caller would require that the mean tag count for a peak be several times larger than the mean tag count in the control. In data such as those used here, this would yield no results, because this level of enrichment simply never occurs. The simple alternative I focused on is to define enrichment as "more tags in the sample than in the control." Of course, one could instead identify the 1% or the 5% of all genes with the largest difference in mean enrichment.

The method proposed here is also not limited to comparing the number of reads in a sample and control experiment. One could use a very similar approach, for example, to compare the log ratio of a sample/control pair to another sample/control pair. This would allow researchers to identify genes that are differently marked for H3K27me3 in cancerous and non-cancerous tissue, for example. A sample and control ChIP-seq would be performed for both the cancerous and the non-cancerous conditions. Rather than identifying the genes that are changed within a single sample/control pair, the goal would be to identify the genes that are marked differently in cancerous and non-cancerous tissue, where each has its own control experiment. In this case, instead of calculating the mean and standard deviation of tag counts in the sample and the control, one could calculate the mean and standard deviation of the log ratios of (sample tags / control tags) for both the cancerous and non-cancerous ChIP-seq data, and test for differences in means in these. As the mean of a variable is normally distributed regardless of the distribution of the underlying variable, the remainder of the test would be unchanged. In the end, one would report genes that are significantly different from one sample/control pair to another.

The method presented here also gets directly to the heart of the original problem: that is, instead of finding the boundaries of enriched regions of unknown length and position throughout the genome, one directly compares control and sample read counts within a single gene. Thus, we greatly simplify both the calculations involved, and in many cases obtain more meaningful information. If the ultimate goal is to identify enriched genes, then identifying the boundaries of enriched regions and translating this into a list of enriched genes can be problematic. In essence, using a peak-caller fails to directly address the actual problem of finding enriched genes or other genomic features. The alternative discussed herein has the potential of greatly simplifying the process of identifying enriched genomic features in ChIP-seq data. This could be especially valuable again in those cases where one is interested in comparing enrichment across sample and control pairs, i.e. across experiments each with their own controls. Identifying enriched regions in each sample relative to their own control, and then looking for differences in the region definition across the two sets of results rapidly becomes intractable. Once the problem is restated to focus on genes or other genomic features, such comparisons become feasible and informative.

As described, the proposed method is a very simple and preliminary approach, and there are clearly many ways in which the statistical framework used here could be improved, some of which are discussed above. Another notable issue is that this method is unable to differentiate between different patterns of enrichment within a bin or gene. Some DNA-binding proteins are known to bind in particular patterns within genes, such as binding specifically in exons but not introns. If the proposed method is used to look only for overall enrichment throughout a gene, this type of pattern will be missed, because the algorithm will only calculate the mean and standard deviation of counts over the entire gene, and will fail to see differences within the gene. Performance in these cases can be improved by using exons, promoters or other genomic features suspected to be particularly enriched for a marker as the binning unit, instead of using genes as shown in this paper. This is a trivial extension of the algorithm, as it simply requires inputting a different set of annotations in place of the gene annotations. In general, it is important to remember that the method used here has no resolution *within* a unit of analysis – that is, this method offers no information about what is going on within a gene (if the gene is the

binning unit). If more resolution is desired, to detect different patterns within genes, binning by exons and introns offers a partial solution. However, some types of signal cannot be seen by this method if annotations do not exist to capture them. More generally, the new method is obviously limited by the availability, and more critically, the accuracy, of genome annotations. While this is not generally a problem in *C. elegans*, which is extensively annotated, lack of accurate annotation could be a barrier to using this method with other genomes, particularly vertebrate genomes. Incorrect gene boundaries, if used for this analysis, could dramatically change results.

Further accuracy in detecting robustly enriched genomic features can also be achieved by making use of independent biological replicates. If the same experiment has been repeated more than once, this algorithm could be used to identify enriched genes in all the replicates. Since there is inherent variability between replicates, we expect that only robustly enriched genomic features would appear in multiple replicates. Therefore, a researcher could choose to consider only genes called in all, or a majority of, replicates as enriched. Doing this is much more straightforward with the proposed method as compared to other peak callers, since the algorithm returns a list of features of interest considered enriched, and therefore the results from different biological replicates are directly comparable. In the case of other peak callers, the edges of called regions rarely coincide across replicates, making reconciling the results of two different biological replicates difficult.

While the method proposed here seems promising for the type of data and problem described in this paper, it would be useful to have further confirmation of the results using ChIP-qPCR or a related method. I presented some preliminary results using RNA-seq data, and found that my method calls more genes with low mRNA levels and calls few highly expressed genes. SICER, on the other hand, calls approximately equal numbers of genes with low and high mRNA levels. Since H3K27me3 preferentially marks genes that are inactive or with low expression levels, these results suggest that my method calls genes enriched for H3K27me3 more accurately than SICER does. Of course, additional confirmation of the peak-calling results would be necessary to reach more definitive conclusions.

Overall, visually, the proposed method of calling enriched genes using the H3K27me3 data yields results that are comparable to, or even better than, those of other peak callers investigated in this paper. Moreover, the results are more immediately useful to researchers attempting to find enriched genes for downstream analysis. In summary, I believe this simple method is worth refining as it may prove a useful tool for researchers attempting to find enriched genes or other genomic features using ChIP-seq data with low signal or high noise.

MATERIALS AND METHODS

I obtained H3K27me3 data from ChIP-seq experiments performed in day 4 adult *Caenorhabditis elegans* worms (Zoey Ni, unpublished data). In these experiments, the sample was prepared by performing ChIP-seq with an antibody targeting H3K27me3, while a control was prepared by performing another ChIP-seq with an antibody targeting the H3 subunit in general. Data were obtained already aligned to the ce6 *C. elegans* reference genome, in the .tag file format. In this format, there are three fields: the chromosome, position and strand (or direction) to which each tag was mapped. The alignment software removed all duplicate reads, which are likely artifacts introduced during PCR amplification, before I obtained the data.

Control Experiment

The ideal control in a ChIP-seq experiment has yet to be determined. Though not the purpose of this paper, a brief summary of current knowledge with respect to ChIP-seq controls is presented here. Many parts of the ChIP-seq experiment can produce artifacts. For example, shearing is not uniform because open chromatin regions tend to be fragmented more easily than closed regions (Kidder et al. 2011). Other sources of artifacts include unrecognized antibody cross-reactivity and variable sequencing efficiency (Kidder et al. 2011). Therefore, using a proper control experiment is vital in order to identify and remove these artifacts from downstream analysis (Park 2009). However, there is no consensus as to how to properly control for these different sources of error. The most commonly used control is input, where a small amount of the DNA present before performing the ChIP is removed and immediately sequenced (Young et al.

2011, Rada-Iglesias et al. 2011, Maruyama et al. 2011). However, other control experiments have been used, including ChIP-seq using a non-specific antibody such as rabbit IgG (Corbo et al. 2010). A more recent evaluation of the use of controls in ChIP-seq experiments concluded that IgG is less desirable because this antibody tends to precipitate much less DNA, which during the amplification step causes overamplification of some regions and poor coverage of others (Kidder et al. 2011). The reasoning behind using H3 as a control is simply that since an entire second ChIP-seq is performed in this case, the H3 experiment can be used to control for both fragmentation bias and artifacts introduced during the ChIP.

Visualizing the data

The data were visualized using the Integrative Genomics Viewer (IGV) (Robinson et al. 2011). For each pair of sample and control tag files, the genome was parsed into non-overlapping 100bp length bins and I counted the number of tags whose putative centers fell in each bin. Since the median length of a tag is about 200bp, the center of a tag was approximated by shifting the start position of the tag by 100bp in the direction of the strand. Since the number of tags in the sample and control are sometimes quite different, the sample tag counts were normalized by multiplying the number of sample tags in each bin by (total # tags in control) / (total # of tags in sample). After tag counts were calculated in each bin for both sample and control, the log enrichment was calculated by taking the log (base 2) ratio of (# sample tags) / (# control tags). Therefore a final value of 1 for a given bin indicates that the bin had twice as many tags in the sample than in the control after normalization. These data were compiled into the standard 'wiggle' file format and loaded into IGV for viewing.

Running ChIPDiff

ChIPDiff was compiled and run according to the instructions in the readme file included in the download. ChIPDiff accepts .tag files as input, and so no pre-processing was necessary. The following settings were used: maxIterationNum = 500 (default), minRegionDist = 1000 (default), minFoldChange = 1.2, minP = 0.95 (Default), maxTrainingSeqNum = 10000 (Default). The parameter minFoldChange represents the

minimum fold change between the sample and control that is required before ChIPDiff will consider any 1000bp bin a putatively enriched region, and include it for further analysis under the HMM model. This parameter had to be set quite low because of the low signal in the H3K27me3 data. Using higher values caused ChIPDiff to pick up only a few – if any – sparse peaks.

Running SICER

I formatted the original .tag files into BED files with the first 6 fields of a standard BED file. The 'dir' field, which indicates the strand (direction) of the read, must be the sixth field, but only four fields are needed, so the fourth an fifth fields were filled in using dummy variables set to zero. These are ignored by SICER. An example of this format is:

chr	posStart	posEnd	dummy1	dummy2	dir
chrX	19923	20023	U0	0	+
chrII	1857	1757	U0	0	-
chrIII	97	197	U0	0	+

In addition, information about the *C. elegans* reference genome used to align the reads from the H3K27me3 data (ce6) was added to SICER's library by modifying the file SICER/LIB/GenomeData.py to add the following lines to the file:

ce6_chroms = ['chrI', 'chrII', 'chrIII', 'chrIV', 'chrV', 'chrX', 'chrM'];

ce6_chrom_lengths = {'chrI':15072549, 'chrII':15279557, 'chrIII':13783768, 'chrIV':17493871, 'chrV':20919680, 'chrX':17719012, 'chrM':13794}

and added 'ce6':ce6_chroms to the species_chroms list, and 'ce6':ce6_chrom_lengths to the species_chrom_lengths list as indicated in the SICER instruction manual.

SICER was run using default parameters. In addition, SICER was also run after modifying a single parameter, window size, to 1000bp.

Running ZINBA

ZINBA was downloaded and compiled at the command line as indicated in the download instructions. The ce6 genome build and mappability file were downloaded from the ZINBA site. BED files were generated from the H3K27me3 tag files that were originally

obtained as indicated under *running SICER*. Preprocessing was performed using the generateAlignability() function per the instructions. ZINBA was run on the result with the 'broad' flag set to TRUE (to indicate that the data contain putatively broad regions of enrichment). All other settings were set to default. ZINBA crashed regularly for unknown reasons and so analysis was not completed on all data.

'New method' algorithm overview

All code was written in the STATA statistical package and is available upon request. Initial binning was performed using the two different methods described in the text. I provide further details here.

Per-base-pair method

In the *per-base-pair* binning method, I used a beta distribution to simulate the length of a potential fragment beyond 150bp to 300bp. The parameters alpha and beta of the cumulative Beta distribution I relied on were 1.2 and 2.1, respectively, as this generated draws such that the median read length was approximately 200bp. Note that for a distribution whose support goes from 150 to 300 to be centered at 200, as in our data, it must be that the distribution of fragment lengths is skewed towards smaller lengths, as the beta with these parameters indeed is. The results of random draws from the Beta(1.2, 2.1) distribution then were scaled up by 151, which is the desired support of the distribution (i.e. = [maximum fragment length (300) and the minimum fragment length (150) + 1]. In other words, a simulated fragment using this method has length

$$fraglen = 150 + \left(rBeta(1.2,2.1) * 151\right)$$

where rBeta is a random draw from the specified Beta distribution. I simulated 1 million fragments to obtain the histogram shown in Fig. 4(B). The probability of a fragment having length less than, or equal to, length L is shown in Fig. 4(A).

To create the per-base-pair counts, I generated a single vector of length 300 containing the values shown in Fig. 4(A), corresponding to the 'contribution' of a read at a given position L past the start position of a fragment. The counts for each base pair were initialized at zero in a giant vector. Then, looping over each fragment in the dataset, the 300 length 'contribution' vector representing the counts was superimposed on the per-base-pair counts vector so that its start position was at the tag start position and it was in

the direction of the tag. Then the per-base-pair vector was added to the 'contribution' vector to produce the updated per-base-pair counts. This was repeated until all fragments had been processed.

After obtaining per-base-pair counts, I input the gene boundaries data, and calculated summary statistics (mean and standard deviations of per-base-pair counts) for the sample and control separately, for each gene. These were used to calculate the t-statistic as indicated in the *t-statistic* section below.

Intermediate binning method

For intermediate binning, I created consecutive length L bins throughout the genome, starting with bin 0, which includes base pairs 1 – (L-1), bin 1, for base pairs L – (2L-1), and so on. Putative tag centers were calculated by shifting the start position of each tag by 100 (median fragment length / 2) in the direction of the fragment. The bin to which each fragment belonged was identified simply by calculating (tag center – 1) / binsize and truncating the result. Subtracting 1 ensures that the first bin is the same length as the others. For example, for a bin size of 200, bin 0 represents 1-200bp, bin 1 201-400bp and so on. Not including this (-1) would cause bin 0 to include only 1-199bp (contains 199bp), while bin 1 represents 200-399 (contains 200bp), since a fragment with a center at 200 would fall in bin 1, not bin 0. In summary, a tag whose center is at position 790bp belongs to the trunc((790-1)/200) = 3rd bin. After each tag was assigned to a bin, I simply counted the total number of tags per bin.

After this binning step, the gene boundaries were input and the same formula ((pos – 1) / binsize) was used to calculate the bin into which the gene start position and end position were located. All bins within that range (inclusive) were considered 'part' of that gene. Note that the gene boundaries computed this way are considerably rougher than using the per-base-pair method, which is exact, and that this worsens with increasing bin size. After all bins belonging to a gene were identified, I calculated the mean and standard deviation of the read counts for those bins for both the sample and the control, and generated the t-statistics as indicated below.

t-statistic

After calculating the mean and standard deviation of counts for both the sample and control in each gene using either binning method described above, I estimated a t-statistic for each gene under the assumption of unequal variances as follows:

$$t = \frac{\bar{x} - \bar{y}}{\left(s_x^2 / n_x + s_y^2 / n_y \right)^{1/2}}$$

where the *x* subscript indicates the sample (H3K27me3) and the *y* subscript indicates the control, xbar and ybar are the mean count in the sample and control respectively, *s* stands for standard deviation, and *n* is sample size.

For the per-base-pair method, *n* = length of gene (in base pairs). For the intermediate binning method, *n* = number of bins in the gene. In the latter case, if *n* = 1 the test could not be performed and the gene was ignored. Because we cannot assume equal variances from the sample and control, I used Satterthwaite's formula to calculate the number of degrees of freedom appropriate under the assumption of unequal variances:

$$df = \frac{\left(s_x^2 / n_x + s_y^2 / n_y \right)^2}{\frac{\left(s_x^2 / n_x \right)^2}{n_x + 1} + \frac{\left(s_y^2 / n_y \right)^2}{n_y + 1}}$$

RNA-seq data processing and analysis

I obtained RNA-seq data for day 4 adult *Caenorhabitis elegans* worms (Mintie Pu, unpublished data). Data were preprocessed before I obtained them by sorting all genes based on their mRNA expression levels, and dividing the list into three quantiles: highly expressed genes, middling expressed genes, and lowly expressed genes (Xiujuan Wang, unpublished data). The lists of highly expressed and lowly expressed genes were used for analysis in this paper. Genes present in both lists were dropped (likely this occurred due to vague genome annotations – lists were annotated using only WikiGene Name, which can be as enlightening as "aminotransferase"), and all duplicates (e.g. IDs appearing more than once) were dropped so only a single instance remained. After censoring and dropping duplicates, there were 4847 genes in the "low" quantile and 5837

genes in the "high" quantile. These lists were annotated using only the "WikiGene Name" set of annotations when obtained (see Ensembl gene annotations, Flicek *et al*. 2011). In order to determine how many "high" and "low" genes were called by SICER, the SICER results, using a bin size of 200bp, were processed to produce a list of called genes. A gene was considered "called" by SICER if it had a greater than 60% overlap with regions called by SICER. In other words, more than 60% of the length of a gene had to fall in a region called by SICER in order for the whole gene to be called enriched. This produced a list of 3643 genes called by SICER. The overlap between the "low" and "high" gene lists, and the SICER list, was then determined using simple merges in STATA (code available upon request). In order to do the same with genes called by my method, I used the list of genes called by the intermediate binning method, with a bin size of 200. The refseq IDs used in the genome annotations in my method were translated into WikiGene Name annotations using Ensembl (Flicek *et al*. 2011), producing 2969 unique IDs. Overlap with the "high" and "low" gene lists was determined in the same way as with the SICER called genes, as was overlap between SICER-called genes and genes called by my method.

REFERENCES

Akkers, R.C., van Heeringen, S.J., Jacobi, U.G., Janssen-Megens, E.M., Françoijs, K-J., Stunnenberg, H.G., Veenstra, G.C. (2009). A hierarchy of H3K4me3 and H3K27me3 acquisition in spatial gene regulation in *Xenopus* embryos. *Dev Cell* **17**(3): 425-434.

Corbo, J. C., Lawrence, K. A., Karlstetter, M., Myers, C. A., Abdelaziz, M., Dirkes, W., Weigelt, M. S., Benes, V., Fritsche, L. G., Weber, B. H. F., and Langmann, T. CRX ChIP-seq reveals the *cis*-regulatory architecture of mouse photoreceptors. (2010). *Genome Res* **20**: 1512-1525.

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Brent, S., Chen, Y., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gordon, L., Hendrix, M., Hourlier, T., Johnson, N., Kähäri, A., Keefe, D., Keenan, S., Kinsella, R., Kokocinski, F., Kulesha, E., Larsson, P., Longden, I., McLaren, W., Overduin, B., Pritchard, B., Riat, H. S., Rios, D., Ritchie, G. R. S., Ruffier, M., Schuster, M., Sobral, D., Spudich, G., Tang, Y. A., Trevanion, S., Vandrovcova, J., Vilella, A. J., White, S., Wilder, S. P., Zadissa, A., Zamora, J., Aken, B. L., Birney, E., Cunningham, F., Dunham, I., Durbin, R., Fernández-Suárez, X. M., Herrero, J., Hubbard, T. J. P., Parker, A., Proctor, G., Vogel, J., Searle, S. M. J. (2011). Ensembl 2011. *Nucleic Acids Research* **39** (suppl 1): D800-D806.

Gottesfeld, J. M. and Melton, D. A. (1978). The length of nucleosomes-associated DNA is the same in both transcribed and nontranscribed regions of chromatin. *Nature* **273**: 317-319.

Maruyama, R., Choudhury, S., Kowalczyk, A., Bessarabova, M., Beresford-Smith, B., Conway, T., Kaspi, A., Wu, Z., Nikolskaya, T., Merino, V.F., Lo, P-K., Liu, X.S., Nikolsky, Y., Sukumar, S., Haviv, I., Polyak, K. (2011). Epigenetic regulation of cell-type specific expression patterns in the human mammary epithelium. *PLoS Genetics* **7**:4.

Park, J.P. (2009). ChIP-seq: advantages and challenges of a maturing technology. *Nature Reviews* **10**, 1038.

Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S.A., Flynn, R.A., Wysocka, J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**:279-285.

Rashid, N.U., Giresi, P.G., Ibrahim, J.G., Sun, W., Lieb, J.D. (2011). ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. *Genome Biology*.

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., Mesirov, J.P. (2011). **Integrative Genomics Viewer**. *Nature Biotechnology* 29, 24–26.

Rozowsky, J., Euskirchen, G., Auerbach, R.K., Zhang, Z.D., Gibson, T., Bjornson, R., Carriero, N., Snyder, M., Gerstein, M.B. (2008). PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nature Biotechnology* **27**(1): 66-75.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin* **2**(6): 110–114.

Wilbanks, E.G., Facciotti, M.T. Evaluation of algorithm performance in ChIP-seq peak detection. *PLoS ONE* **5**(7): (2010).

Young, M.D., Willson, T.A., Wakefield, M.J., Trounson, E., Hilton, D.J., Blewitt, M.E., Oshlack, A., Majewski, I.J. (2011). ChIP-seq analysis reveals distinct H3K27me3 profiles that correlate with transcriptional activity. *Nucleic Acids Research*.

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Meyers, R. M., Brown, M., Li, W. and Liu, X. S. (2008). Model-based Analysis of ChIP-Seq (MACS). *Genome Biol* **9**(9): R137.

Zang, C., Schones, D.E., Zeng, C., Cui, K., Zhao, K., Peng, W. (2009). A clustering approach for identification of enriched domains from histone modification ChIP-seq data. *Bioinformatics* **25**: 15.

FIGURES

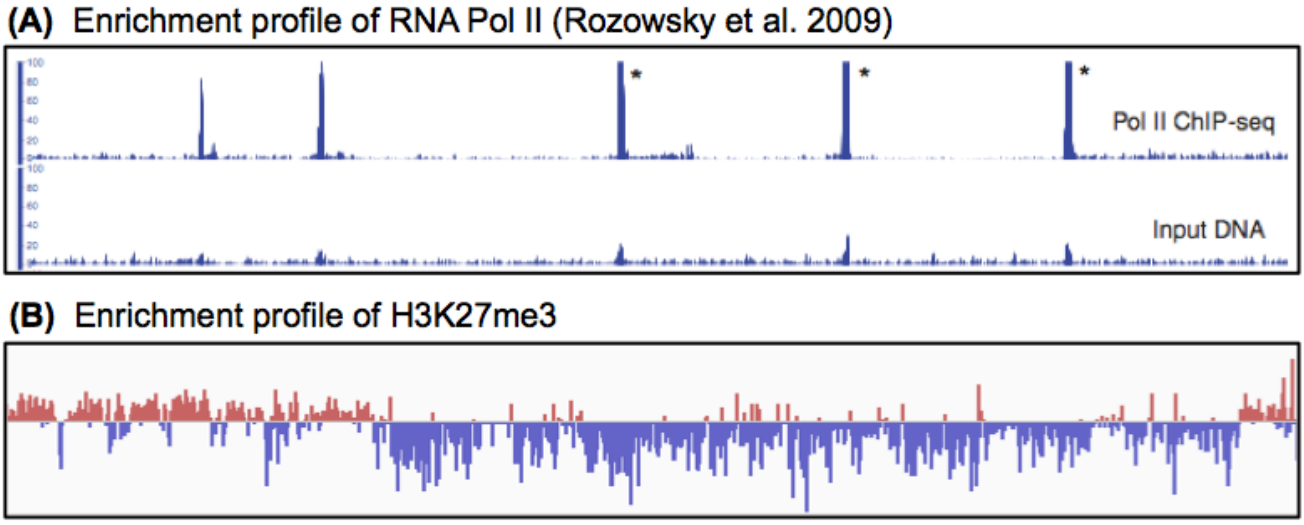


Fig. 1: Comparison of 'sharp' (A) and 'broad' (B) ChIP-seq enrichment profiles. Pol II sample (top track) and control (input DNA, bottom track) are shown here separately and the y-axis is the number of reads (0-100), while the H3K27me3 data shown are the log(base 2) of (# sample tags / # control tags) (axis values not shown, ranging from approx. -1 to +1). In (B), positive (red) values represent enrichment of sample relative to control, while blue (negative) signifies the opposite. Pol II data show dramatic, narrow peaks of sample tags while the H3K27me3 data do not.

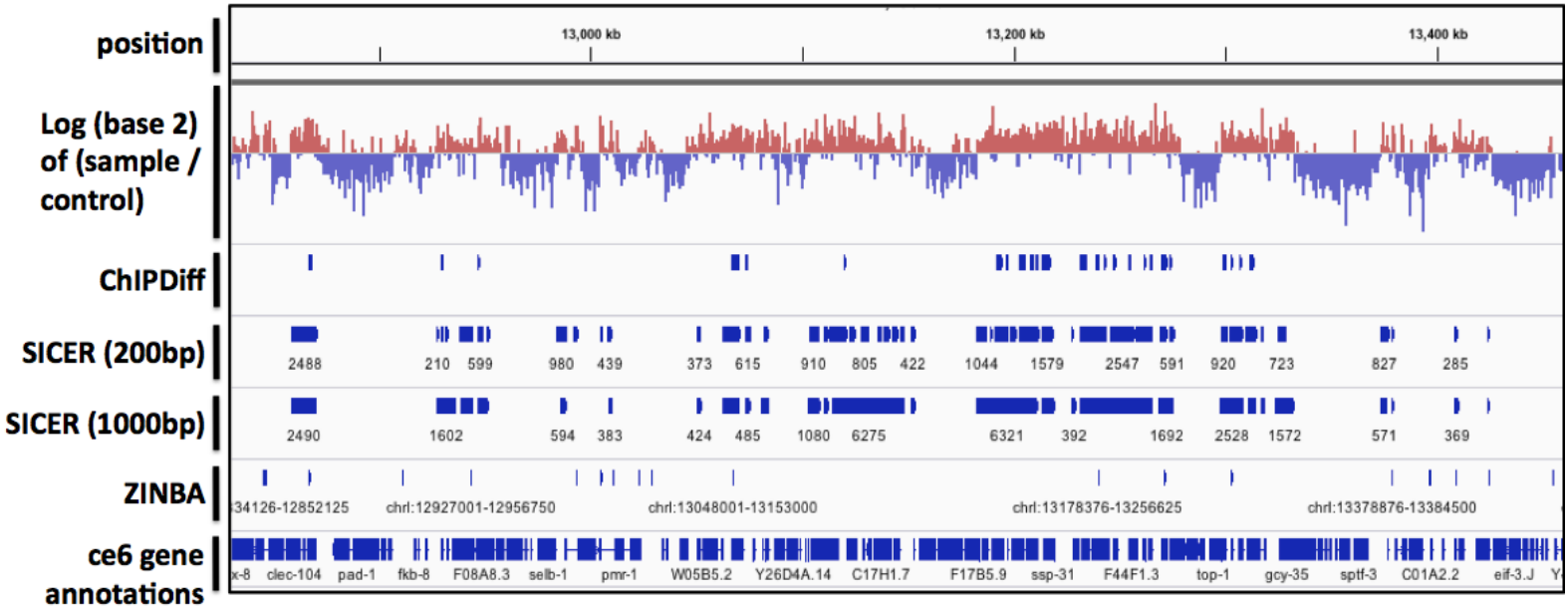


Fig. 2: H3K27me3 data were plotted in IGV using the log(base 2) calculated as described in the methods. Positive values (red) indicate enrichment of reads in the sample relative to the control. The regions determined to be enriched for the sample by ChIPDiff, SICER using 200bp windows, SICER using 1000bp windows, and ZINBA are shown. For reference, the ce6 gene annotations are also shown in the IGV 'compressed' format. Region shown is from approximately 13,020kb-13,480kb on chrI.

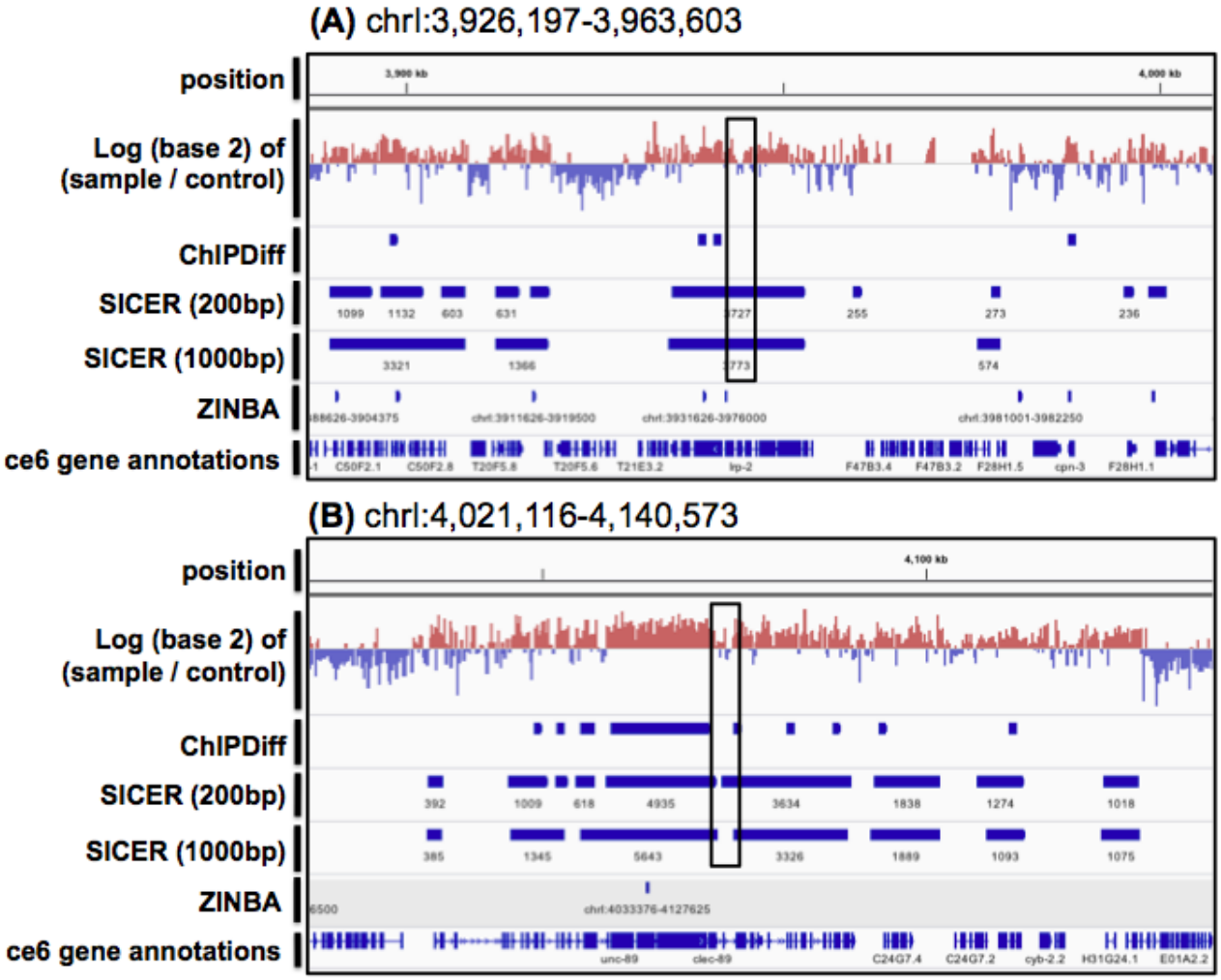


Fig. 3: Both (A) and (B) are shown on the same scale. This figure illustrates the variability in the definition of an enriched region by SICER. In (A), a small negative (blue) region (in black box) does not interrupt an 'enriched' region called by SICER. In (B), a very similar negative region causes a gap between two regions.

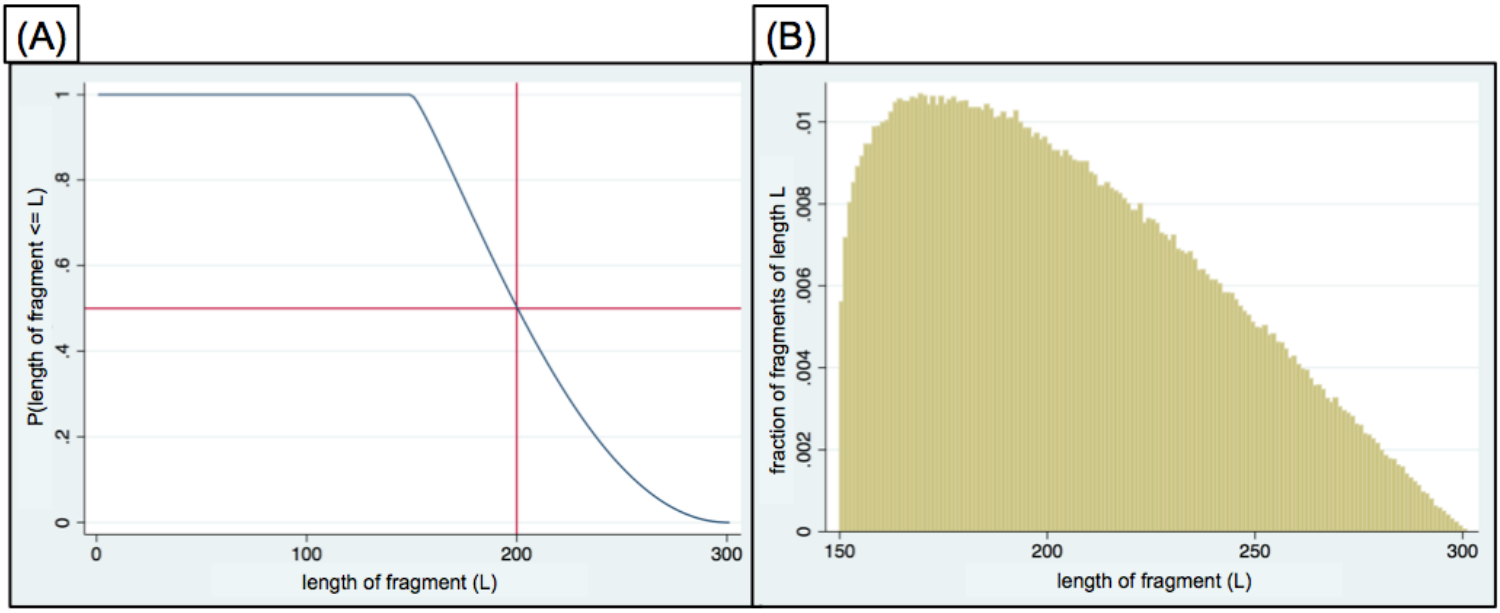


Fig. 4: (A) Graph showing the contribution of a single fragment starting at position 1 to the per-base-pair counts. A fragment is assumed to be between 150 – 300bp, with a median length of 200bp. Therefore a fragment contributes 1 to the per-bp counts for the first 150 bps after the start position because it is guaranteed to be present for at least 150 bp, and then contributes progressively less until reaching zero at 300, due to a declining probability that the fragment continues to exist at that length. At 200bps, the contribution of the fragment to the per-bp counts is exactly 0.5, signifying that 200bps represents the median fragment length. (B) Histogram showing the distribution of 1 million fragments simulated using the model proposed in (A). Median fragment length is roughly 200bp, and is skewed towards smaller fragments.

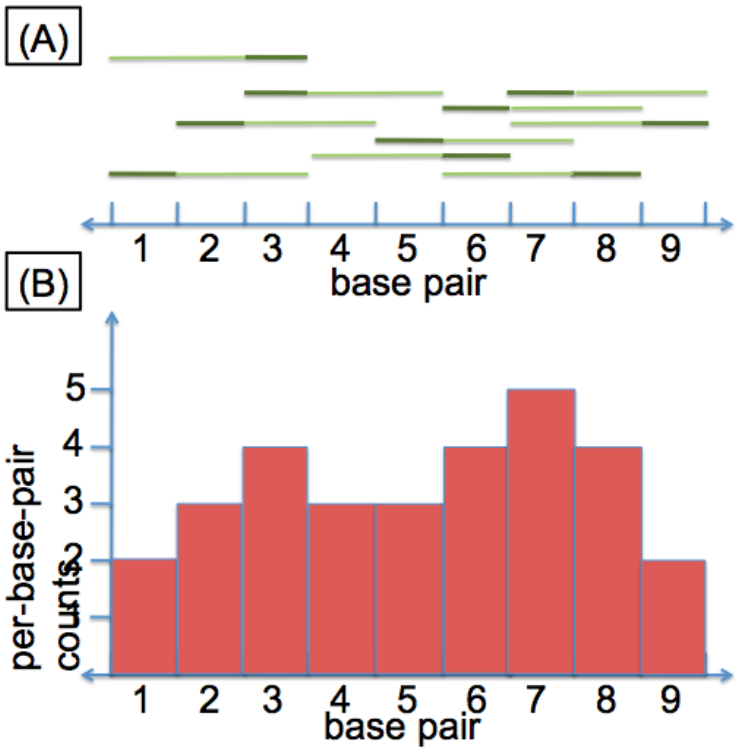


Fig. 5: Illustration of how per-base-pair counts were obtained. In this simple example, tags are 1bp in length (dark green) and the whole read is 3bp in length (extension is light green) as shown in (A). When read lengths are known, the base pair counts are simply the sum of the number of fragments overlapping at each position, as shown in (B).

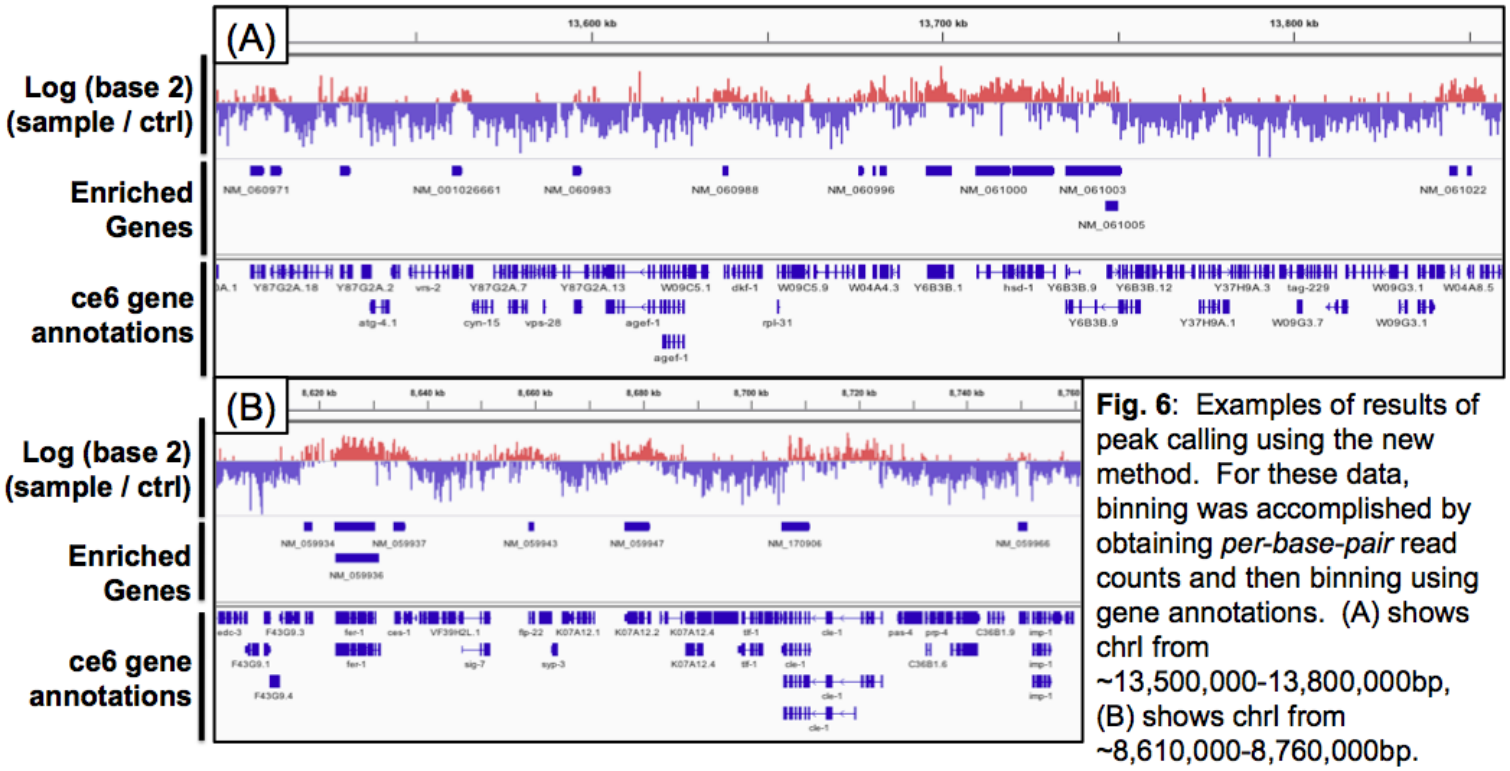


Fig. 6: Examples of results of peak calling using the new method. For these data, binning was accomplished by obtaining *per-base-pair* read counts and then binning using gene annotations. (A) shows chrI from ~13,500,000-13,800,000bp, (B) shows chrI from ~8,610,000-8,760,000bp.

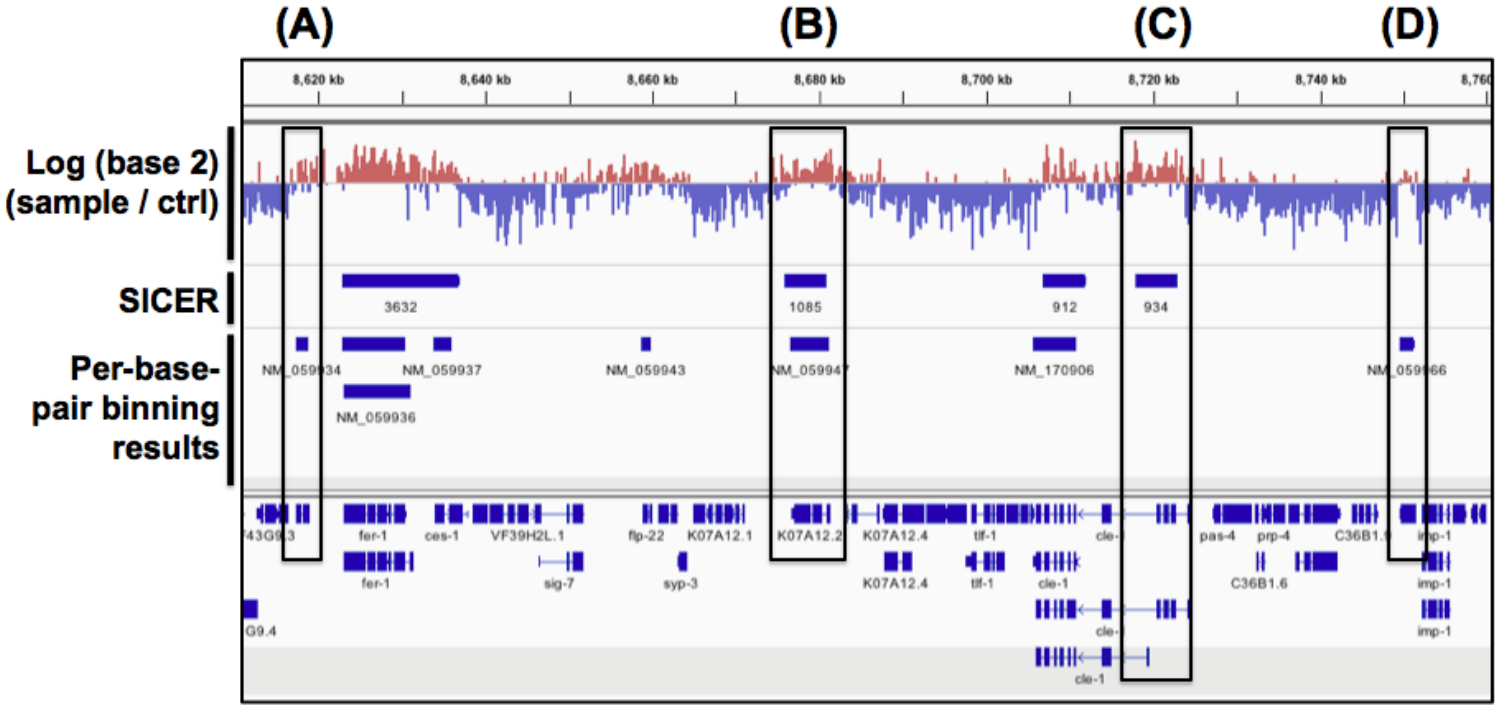


Fig. 7: Comparison of SICER (1000bp windows) and new method using the per-base-pair binning method. The new method was run using gene annotations, looking for enrichment within each gene. In (A) and (D), relatively small genes occurring in small regions of positive enrichment are flagged in the new method, whereas with SICER the regions are too small to be identified. In (B), nearly identical regions were found by both methods. In (C), SICER identified a region but the new method did not, because there are no genes in that location. For a researcher attempting to find genes enriched for this marker, the region identified by SICER in (C) is not informative, since the new method shows that overall the genes overlapping with that region are not enriched.

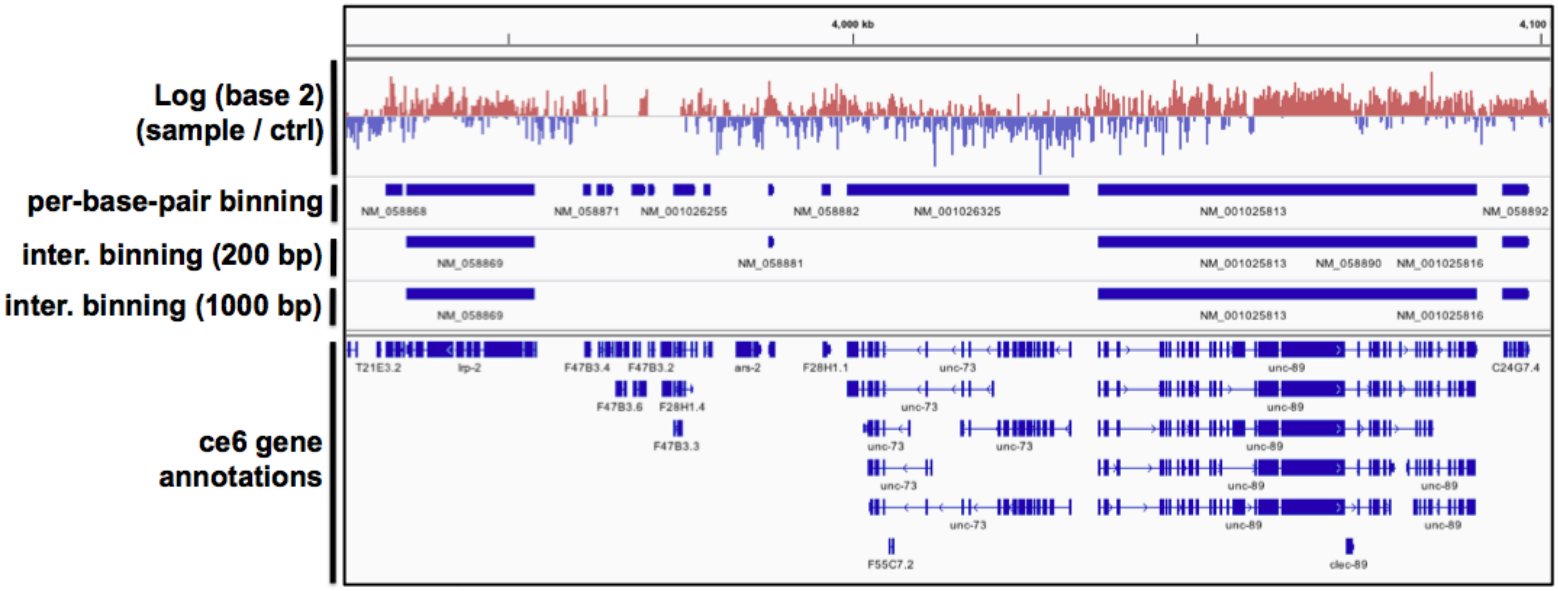


Fig. 8: Comparison of the two binning methods. The intermediary binning method was used with two different bin sizes of 200bp and 1000bp, and a significance level of 0.01.
chrI: 3,958,093-4,100,237

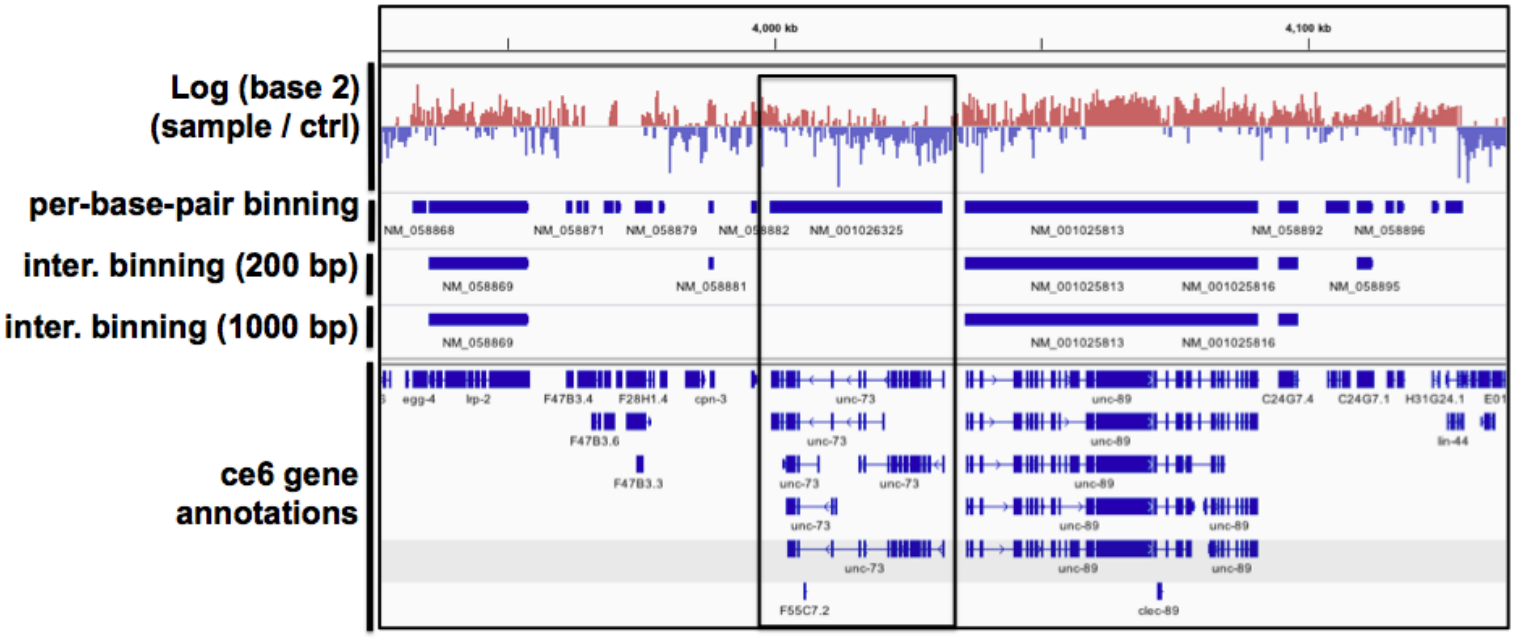


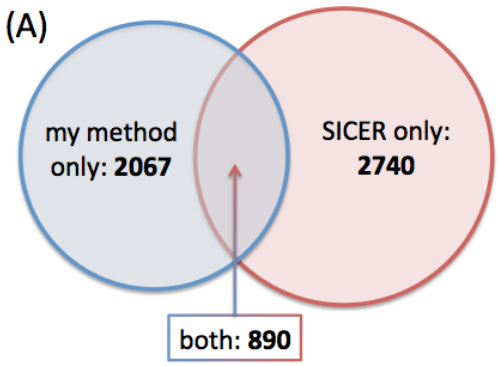
Fig. 9: Comparison of the two binning methods. The intermediary binning method was used with two different bin sizes of 200bp and 1000bp. The boxed region shows a case where the more conservative results of the intermediary binning method are an advantage over the blunt per-base-pair binning approach, since the latter approach identifies a gene that does not visually appear to be enriched.
chrI: 3,921,654-4,102,779

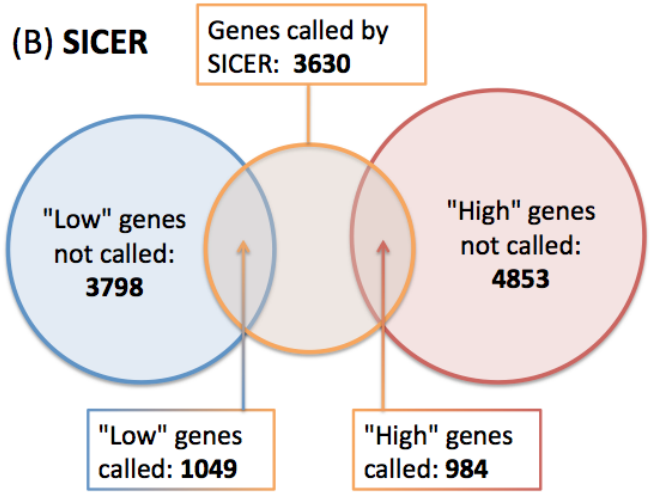
	SICER	My Method	Both	Neither	Total
"LOW" expression	1049	1276	420	2942	4847
"HIGH" expression	984	331	79	4601	5837
Neither	1597	1350	391	0	2556
Total	3630	2957	890	7543	

Table 1A. Raw results of overlap between the RNA-seq "highly expressed" and "lowly expressed" gene lists and the genes called by SICER or my method. For example, 1049 genes were called by SICER and considered to have low expression levels in the RNA-seq data, and 4847 genes were considered to have low expression regardless of whether they were called or not. Results show that my method calls proportionally more "LOW" genes and fewer "HIGH" genes than SICER.

	Proportion of Genes that are "LOW"	Proportion of Genes that are "HIGH"	Proportion of Genes that are neither
SICER	0.2890	0.2710	0.4400
My method	0.4315	0.1119	0.4566

Table 1B: Proportion of genes called by SICER and my method that are highly expressed, lowly expressed, or neither in the RNA-seq data. For example, since SICER calls a total of 3630 genes and 1049 of them are lowly expressed in the RNA-seq data, the proportion of genes that are "LOW" for SICER is 1049/3630 = 0.2890. Results show that my method calls proportionally more "LOW" genes and fewer "HIGH" genes than SICER.





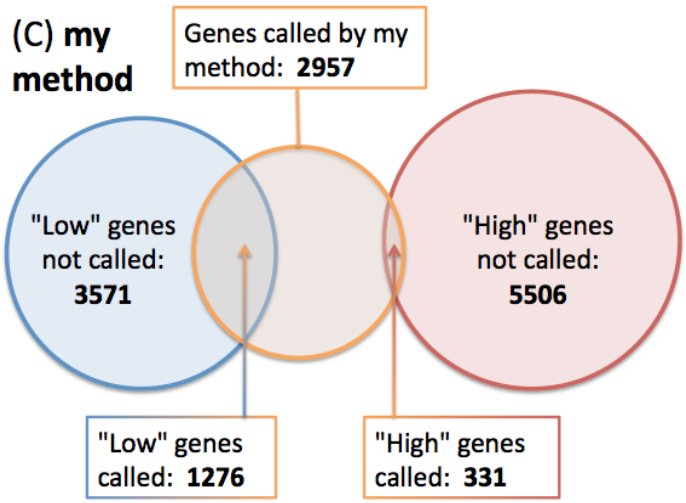
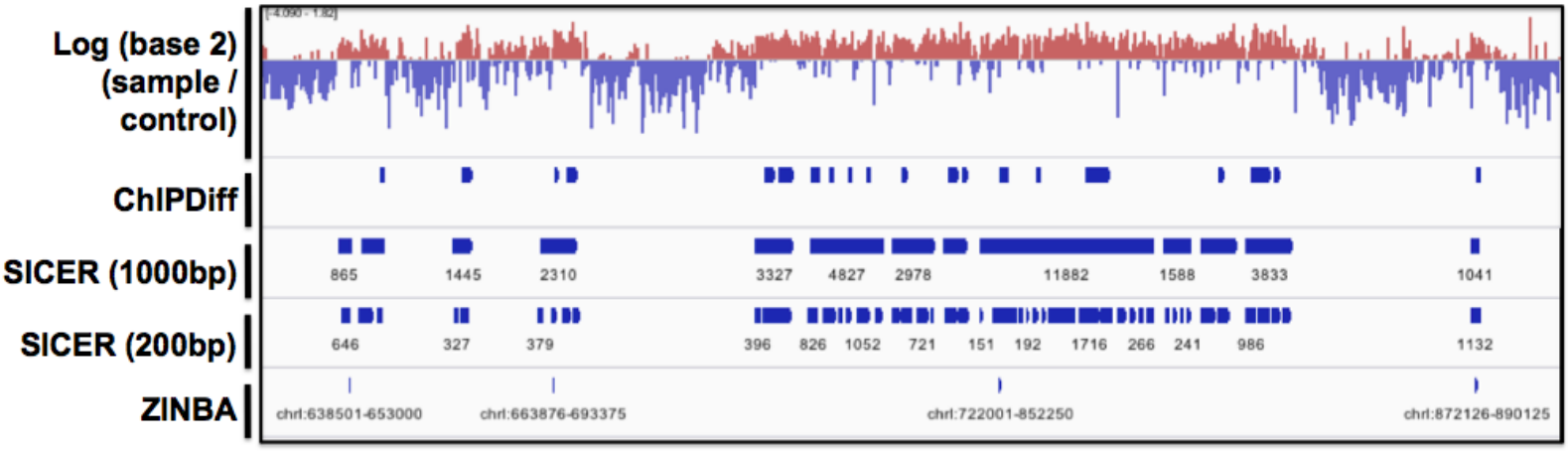
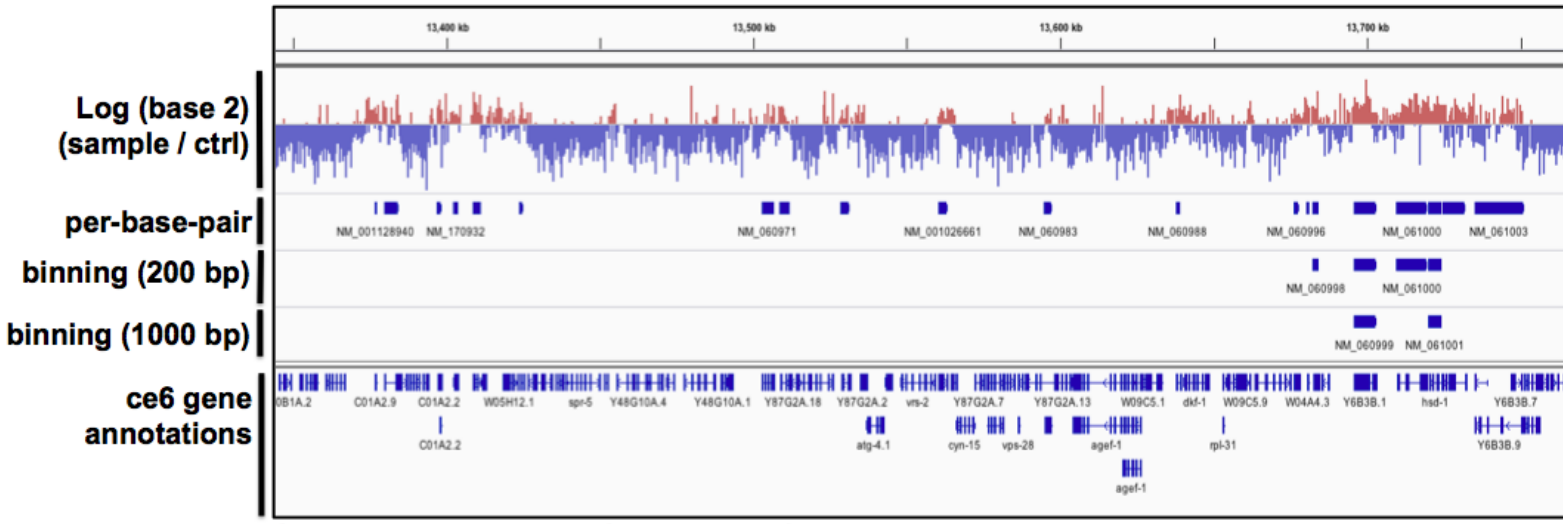


Fig. 10: (A) Overlap between genes called by SICER and by my method (intermediate binning, bin size 200). (B) Overlap between "low" expressed genes from RNA-seq data, "high" expressed genes, and SICER results. (C) Overlap between "low" expressed genes from RNA-seq data, "high" expressed genes, and results of my method with intermediate binning (bin size 200). While SICER calls similar numbers of "high" and "low" expressed genes, my method calls many more "low" and fewer "high," which is more consistent with what we would expect to see for genes enriched with H3K27me3.

SUPPLEMENTARY FIGURES



Supplemental Fig. 1: H3K27me3 data were plotted in IGV using the log(base 2) values as described in the methods. Positive values (red) indicate enrichment of reads in the sample relative to the control. The regions determined to be enriched for the sample by ChIPDiff, SICER using 200bp windows, SICER using 1000bp windows, and ZINBA are shown. Region shown is chrI: 445,828-1,044,350 bp.



Supplementary Fig. 2: Comparison of the two binning methods. The intermediary binning method was used with two different bin sizes of 200bp and 1000bp. chrI: 13,342,755-13,759,601